A Priori Justification for Effect Measures in Single-Case Experimental Designs

Rumen Manolov[1], Mariola Moeyaert[2], & Joelle E. Fingerhut[2]

[1]Department of Social Psychology and Quantitative Psychology,

Faculty of Psychology, University of Barcelona

[2] Division of Educational Psychology & Methodology, Department of Educational and Counseling Psychology,

School of Education, University at Albany

***Corresponding Author:**
Rumen Manolov*

Department de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona

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

E-mail: rrumenovl3@ub.edu

Mariola Moeyaert

Division of Educational Psychology & Methodology, Department of Educational and Counseling Psychology, School of Education, University at Albany

Office Catskill 271. 1400 Washington Avenue, Albany, New York 12222, USA

E-mail: mmoeyaert@albany.edu

Joelle E. Fingerhut

Division of Educational Psychology & Methodology, Department of Educational and Counseling Psychology, School of Education, University at Albany

1400 Washington Avenue, Albany, New York 12222, USA

E-mail: jfingerhut@albany.edu

Declarations

Funding: No funding was received for the current text

Conflicts of interest/Competing interests: The authors report no conflicts of interest

Availability of data and material: No data have been gathered or re-analyzed in the context of the current manuscript. Nevertheless, there is supplementary material that can be consulted in Appendix A (https://osf.io/t96fc/): it includes quotes from several methodological and statistical articles presenting or discussing specific quantitative data analysis approaches.

Code availability (software application or custom code): Several freely-available software applications are mentioned in Appendix B (https://osf.io/t96fc/), but the underlying code for creating these application is not been publicly shared in the context of the current manuscript.

Abstract

Due to the complex nature of single-case experimental design data, numerous effect measures are available to quantify and evaluate the effectiveness of an intervention. An inappropriate choice of the effect measure can result in a misrepresentation of the intervention effectiveness and this can have far-reaching implications for theory, practice and policymaking. As guidelines for reporting appropriate justification for selecting an effect measure are missing, the first aim is to identify the relevant dimensions for effect measure selection and justification prior to data gathering. The second aim is to use these dimensions to construct a user-friendly flowchart or decision tree guiding applied researchers in this process. The use of the flowchart is illustrated in the context of a preregistered protocol. This study is the first study that attempts to propose reporting guidelines to justify the effect measure choice, before collecting the data, to avoid selective reporting of the largest quantifications of an effect. A proper justification, less prone to confirmation bias, and transparent and explicit reporting can enhance the credibility of the single-case design study findings.

Keywords: single-case experimental design; statistical analysis; quantitative methods; reporting standards; scientific rigor

A Priori Justification for Effect Measures in Single-Case Experimental Design

Single-case experimental designs (SCEDs) offer the possibility to gather data repeatedly under different conditions, manipulated actively by the researchers (Horner et al., 2005). The aim is to obtain evidence regarding the effectiveness of the intervention for the single participant or the few participants studied. The usefulness of SCED studies for providing strong evidence is boosted by meeting the criteria of methodological rigor (Ganz & Ayres, 2018), whereas drawing more general conclusions requires replicating the results in several studies using the same intervention for the same problematic aspect (Kennedy, 2005; What Works Clearinghouse, 2020).

In terms of how to analyze SCED data, visual analysis is commonly considered as a first step, especially for performing a formative analysis during data collection, whereas quantitative techniques are useful for summative analysis post data collection (Ledford et al., 2019). Regarding quantitative analysis, there is a lack of consensus about which techniques are most appropriate (Busse et al., 2015; Smith, 2012). One option would be to seek guidance from methodological quality scales, but they rarely include items rating the quality of the quantitative data analysis technique. These scales do not go beyond visual analysis, and for the assessment of social validity (Ganz & Ayres, 2018; Lobo et al., 2017; Maggin et al., 2014; Wendt & Miller, 2012). One of the scales, called “Risk of Bias in N-of-1 Trials” (Tate et al., 2015), however, puts the emphasis on the justification provided for choosing one of the available quantifications, but there are no (reporting) guidelines for appropriate justifications. A second option would be to consult the recommendations made in textbooks dedicated to applied SCED research. In texts explaining the use of SCED in different contexts, there have been different approaches to dealing with the choice of a data analytical approach. On the one hand, there have been

recommendations to prioritize visual analysis over statistical or quantitative analysis (Janosky et al., 2009; Kennedy, 2005; Riley-Tillman et al., 2020). On the other hand, there have been reviews of multiple quantitative options with an emphasis on the importance of their assumptions and the data features they quantify (Moeyaert et al., 2018; Tate & Perdices, 2019). As intermediate options, there has been an emphasis on "readily available" (p. 162) or "user-friendly" (p. 165) techniques (Barker et al., 2011), or on descriptive statistics (Janosky et al., 2009). Specific techniques such as nonoverlap indices and randomization tests have also been recommended due to the lack of parametric assumptions and the ease of understanding them and computing them with the available software tools (Morley, 2018). In summary, detailed guidelines for selecting effect measures are missing from textbooks presenting SCEDs to applied researchers from different fields (e.g., special education, clinical psychology, sport psychology, neurorehabilitation, biomedicine). In the current text, we will provide guidance, on the basis of methodological and statistical texts.

A Note on Terminology

When referring to the analysis of SCED data beyond visual inspection, a potentially more inclusive term could be “quantitative analysis techniques”, whereas a more restrictive term would be “effect size measures”. For instance, a randomization tests, which can be conceptualized as tools for statistical inference, would be a quantitative analysis technique that can be applied to different descriptive effect size measures (e.g., a nonoverlap index or a mean difference, Heyvaert & Onghena, 2014). As another example of a “quantitative analysis techniques” that includes an “effect size measure”, multilevel models can be mentioned. Multilevel models are modeling options that can be implemented via different estimation procedures (e.g., restricted maximum likelihood and Bayesian; Moeyaert et al., 2017) and can be

used for estimating different effect size measures (e.g., a mean difference or a difference in slope).When focusing on descriptive measures of effect, most can be understood as “effect size measures”, but for nonoverlap indices it has been put into doubt whether they actually quantify the *magnitude* of effect (Carter, 2013; Pustejovsky, 2018; Solomon et al., 2015; Wolery et al., 2010). Thus, nonoverlap indices can be distinguished from “effect size measures” (e.g., a raw or a standardized mean difference) in which not only the ordinal superiority of conditions is quantified, but also the distance between conditions or the degree to which the conditions are different (Natesan Batley et al., 2020). Thus, in the current text we use the term “effect measures” with which we aim to include both nonoverlap indices and “effect size measures”. When necessary, we also refer to randomization tests (a technique for design-based inference) and to HLMs making possible model-based inference (Onghena, 2020) for descriptive effect measures or effect size measures.

Aim and Organization of the Text

Ambiguous or unreported choices in relation to selection of an effect measure and the data analytical plan in general (including selective reporting, i.e., only reporting the effect measures that are well-aligned with the researchers’ hypotheses) can be considered an example of questionable research practices that can lead to biased results (Hantula, 2019). This is relevant because the choice of effect measures may affect the conclusions regarding intervention effectiveness (Simmons et al., 2011). For instance, Beckers et al. (2020) performed a review of SCED research in children with cerebral palsy and reported that many studies conducted statistical analysis, but justification was missing. The complication resides in the fact that statistical analysis can involve multiple approaches and different effect measures; therefore, justifying the use of statistical analysis, in general, does not necessarily provide information

about the reason for selecting a specific effect measure. In addition, wrong interpretations of estimated effects might be provided if researchers do not have a good conceptual understanding of the effect measure to begin with.

Multiple dimensions can be considered and evaluated simultaneously, when selecting an effect measure. An initial goal of the current text is to identify these dimensions, on the basis of SCED literature (i.e., recommendations from methodological and statistical articles and applied research articles). On the basis of these dimensions and their facets, a second aim is to provide a user-friendly flowchart in order to guide applied researchers providing an appropriate justification for their effect measure selection. The explicit justification of the effect measure selected is expected to improve reporting by enhancing the transparency of the decision process.

It is necessary to highlight that we will only briefly mention, but not discuss in detail, visual analysis (Lane & Gast, 2014; Ledford et al., 2019; Maggin et al., 2018) or masked visual analysis (Byun et al., 2017; Ferron et al., 2017), which do not lead to a quantification of the magnitude of the intervention effect. The aim is also not to present in detail the benefits and pitfalls of different quantitative analysis techniques, as such information is already available elsewhere (Busse et al., 2015; Chen et al., 2015; Gage & Lewis, 2013; Lobo et al., 2017; Manolov & Moeyaert, 2017a, 2017b; Solomon et al., 2015).

First, in the following sections, the importance for justifying a priori the selection of a quantitative analysis technique in general (incl. an effect measure) is presented. Second, a brief overview of quantitative techniques is provided together with their justified use (provided by the founders or developers of these techniques). Third, the main dimensions for justifying the selection of an effect measure are discussed and organized in a flowchart. SCED researchers are encouraged to use the flowchart in their future studies as a guidance to justify their choice prior

to the start of data collection. Finally, in order to illustrate how to use the flowchart as part of the a priori data analytic plan, an empirical example is provided.

A Priori Aspects to Include in the Data Analytic Plan

Several decisions are required regarding data analysis before the data collection ends (De Young & Bottera, 2018) or, even more appropriately, before data collection begins. These decisions can be made explicit either as part of a preregistered protocol or as part of the data analysis section of the article presenting the results of an empirical study. In order to enhance transparency and avoid experimenter biases, researchers are highly encouraged to make the data analytical plan publicly available to the broader research community. This can be done through submitting the research protocol to a journal or to the Open Science Framework (OSF; https://osf.io/). In that way the study protocol, containing the data analytic plan, is registered prior to starting data collection (Hales et al., 2019; Johnson & Cook, 2019). Preregistered protocols are a methodological safeguard against confirmation bias relevant for science in general (Nuzzo, 2015). Such protocols have been recently advocated for by study authors in the field of psychology (e.g., Gonzales & Cunningham, 2015; Nosek et al., 2018), psychopathology (Krypotos et al., 2019) and rehabilitation (Krasny-Pacini & Evans, 2018), as well as by journal editors (Jonas & Cesario, 2016; Lindsay, 2015) and institutions (e.g., Institute of Education Sciences, 2020). Preregistration has also been emphasized recently in the SCED context (Johnson & Cook, 2019). To enhance this practice, tools have been made freely available online (e.g., https://osf.io/zab38/, https://cos.io/prereg/).

In the context of N-of-1 trials (which could be understood as a specific kind of SCED more common in medical research; Nikles & Mitchell, 2015, Tate & Perdices, 2019), it has been recommended that “all statistical methods planned—from visual representation to meta-

analysis—should be described in the protocol" (Porcino et al., 2020, p. 10), including effect sizes, statistical significance, ways of performing sensitivity analyses, and how heterogeneity between participants will be assessed. In the SCED context, the importance of explicitly describing the expected data pattern and the expected effect of the intervention has also been recently emphasized (Maggin et al., 2020).

In the SCED context, the variety of effect measures available, and the lack of consensus regarding the optimal one (Busse et al., 2015), has led to the recommendation to report a variety of different effect measures (Vannest et al., 2018). This enables assessing the consistency in findings related to intervention effectiveness. If the same results are obtained regardless of the chosen effect measure, then researchers can be more confident in making statements about intervention effectiveness (Lobo et al., 2017). However, it may also lead to finding *at least one* effect measure providing evidence in support of an effective intervention. As Kratochwill et al. (2018) state "selective results may also appear in cases where multiple-outcome measures are included in a single investigation" (p. 71). Thus, an unwanted side effect of the recommendation of applying several effect measures is that researchers might only report the one that gives evidence in support of the intervention (i.e., selective reporting; Vannest et al., 2018). An informed selection of an effect measure may reduce the probability of reporting on such spurious findings. As Levin et al. (2017, p. 29) state, "if the researcher does not specify a particular anticipated effect type on an a priori basis (and, particularly, prior to examining the data), but rather conducts multiple analyses on the same data with different effect-type specifications, then we would again have ethical concerns and would question the validity of the researcher's statistical conclusions."

Overview of Single-Case Quantitative Analysis Techniques

In order to provide guidance on what to include as proper justification for selecting a quantitative analysis technique, a good conceptual understanding of the different alternatives is needed, together with a good understanding of their intended use as stated by their founders. For that purpose, a document was created (Appendix A, available at https://osf.io/t96fc/?view_only=19c595b82d9c4a96aec04eeaf6a4b196) including quotes from the founders of effect measures regarding their main features and uses.

Before presenting a brief review of effect measures, a remark is needed on replication and randomization, which are two key features of SCEDs relevant for their internal and external validity (Horner et al., 2005; Kratochwill & Levin, 2010).

Replication

The basic effect (i.e., an A-B comparison between a baseline and an intervention phase) is the building block for quantitative data analysis, but this basic effect needs to be replicated in order to have greater confidence that the effect is due to the intervention. Most effect measures (e.g., nonoverlap indices and log-response ratio) have been initially proposed and discussed for the quantification of a basic effect, although others (e.g., hierarchical linear models and design-comparable effect size) are especially developed for combining effects. Thus, replication is not only a necessity for internal and external validity, but it also informs the unit of analysis. For designs that entail a replication within the participant (e.g., withdrawal/reversal designs, alternating treatment designs [ATDs], and changing criterion designs [CCDs]), the unit of analysis is the participant and we refer to as “within-case” effect measures. For designs that include replication across participants (e.g., multiple-baseline designs), the unit of analysis can be the participants and/or the study. For the latter case, the term “across-case” effect measures.

Randomization and Randomization Tests

Randomization (i.e., random assignment of measurement times to conditions or random choice of the moments of change in condition) is a design feature that makes it possible to use randomization tests as an analytical option. Randomization tests do not require parametric data assumptions and are applicable even when there are missing data (De et al., 2020). Randomization tests allow flexibility in defining the test statistic (Heyvaert & Onghena, 2014). Therefore, the decision to use a randomization test does not determine the effect measure to be used, as a randomization test can be applied to multiple data features such as level, trend, variability, overlap and immediacy (Tanious et al., 2019). Specifically, when focusing on an overall difference in level, a mean difference can be used as an effect size measure (e.g., Ferron & Ware, 1995). It also possible to define the test statistic according to whether the change in level is expected to be immediate, comparing the last three baseline phase measurements and the first three intervention phase measurements (Michiels & Onghena, 2019), or delayed, excluding the initial values of the intervention phase (Levin et al., 2017). Alternatively, focusing on trend, the difference between slopes can be used as an effect size measure, whereas focusing on variability, the difference between conditions can be quantified via a variance difference or a variance ratio (Levin et al., 2020). In terms of overlap, the Nonoverlap of All Pairs (Parker & Vannest, 2009) can be used as a test statistic and effect measure (Heyvaert & Onghena, 2014a). It is also possible to use consistency measures as test statistics (Tanious et al., 2020). Furthermore, specific proposals for test statistics have been made for ATDs (Manolov & Onghena, 2018, and Manolov, 2019, suggest comparing the data paths represented by the lines connecting the measurements from the same condition) and for CCDs (Onghena et al., 2019, suggest using the mean absolute deviation between the measurements and the criteria). Finally,

when using a randomization test, apart from choosing an effect measure, it is important to select a randomization scheme that is appropriate for the specific SCED (see Levin et al., 2018, for multiple-baseline designs; Onghena, 1992 for withdrawal/reversal designs; Levin et al., 2012, and Onghena & Edgington 1994, 2005 for ATDs; and Ferron et al., 2019, and Onghena et al., 2019, and for CCDs).

In summary, a randomization test can be used in conjunction with visual analysis, mean differences, nonoverlap indices, or even with multilevel models (Michiels et al., 2020). Thus, a randomization test can use a within-case or an across-case effect measure as a test statistic. It should be noted that the purpose of using a randomization test is for tentative causal inference (not for population inference) and not for quantifying the magnitude of effect. Randomization can also be used for controlling false positives in the context of response-guided experimentation when performing masked visual analysis (Byun et al., 2017; Ferron et al., 2017; Joo et al., 2018). In this context of use of randomization, the aim is not produce an effect measure. For these reasons, the current text does not discuss randomization tests further, although using a randomization test is recommended whenever there is randomization in the design.

Within-Case Effect Measures

Nonoverlap Indices

Vannest and Ninci (2015) advocate for nonoverlap indices because these are easily calculated by hand and are easily interpreted, and do not require normally distributed data. Nonoverlap indices are especially justified when the data cannot be meaningfully represented by a mean or trend lines (Parker, Vannest, & Davis, 2011). The nonoverlap indices can be used if the sole interest is in quantifying the percentage of data separation between different phases.

Some nonoverlap indices require absence of baseline trend: this is the case, for instance, of the Nonoverlap of All Pairs (NAP; Parker & Vannest, 2009). Other nonoverlap indices control for trend: Tau-U with baseline trend control (Parker, Vannest, Davis, et al., 2011) and the Baseline corrected Tau (Tarlow, 2017). Thus, it has to be emphasized that not all nonoverlap indices have similar features or strengths and limitations (see Parker, Vannest, & Davis, 2011, for a review) and that variants of one index such as the Tau can be quite different. For instance, the Tau-U and Baseline corrected Tau are different in how they control for baseline trend (Manolov, 2018; Tarlow, 2017). Only a limited number of nonoverlap indices have an established sampling distribution (i.e., NAP and Tau-U without baseline control). However, the interpretation of the confidence intervals and *p*-values associated with these indices is subjected to the assumption of independent data (Pustejovsky & Swan, 2018). Moreover, nonoverlap indices do not quantify the magnitude of intervention effectiveness (Carter, 2013); specifically, they cannot quantify differences in amount of separation between data points once complete overlap is achieved.

Regression-Based Quantifications

Simple ordinary least squares (OLS) regression can be used to quantify the change in outcome level between baseline and intervention conditions (or between intervention conditions in the case a baseline phase is missing, as in the case of ATDs in which the relative effectiveness of several interventions is commonly compared). Piecewise regression is an extension of simple OLS, taking into account time trends during the baseline and intervention conditions (Center et al., 1985, Van den Noortgate and Onghena, 2003). Instead of providing a quantification of changes in level, it provides separate quantifications of changes in level and in slope due to the intervention. The quantification can be expressed both in raw and in standardized units (Van den Noortgate & Onghena, 2003, 2008). Another regression-based quantification is obtained in the

context of a generalized least squares (GLS) regression (Swaminathan, Rogers, Horner, et al., 2014), for which a Bayesian approach for drawing inferences has also been presented (Swaminathan, Rogers, & Horner, 2014). The effect size proposed by Swaminathan and colleagues, on the basis of the regression model, is an overall effect, combining the change in level and the change in slope and can be expressed in raw or standardized units. GLS is different from OLS in that it can model several functional forms of autocorrelation and it also enables modeling count outcomes (Swaminathan, Rogers, & Horner, 2014). In terms of the appropriate situations for applying GLS, sufficient data are necessary for estimating autocorrelation precisely and the change in level and in trend need to be visually inspected in order to assess the meaningfulness of the quantification (Maggin et al., 2011).

Log-Response Ratio

The log response ratio has been advocated for on the basis of its insensitivity to procedural details such as series length and observation session length (Pustejovsky, 2019), as well as due to the possibility to express it in meaningful terms as a percentage change (Pustejovsky, 2018). Its use is justified when the intervention does not consistently lead to the extinction of the target behavior and when there are no time trends and autocorrelation (Pustejovsky, 2018).

Comparison to a Pre-Defined Goal

Visually, a goal line has been suggested to be superimposed on the graph with the time-series data (Riley-Tillman et al., 2020). Quantitatively, the number of sessions required to reach a pre-established criterion can be counted (Kipfmiller et al., 2019). In terms of a quantification of

the effect, the percent of goal obtained (Ferron et al., 2020) expresses the level achieved in relation to the pre-established goal.

Across-Case Quantitative Analysis Techniques

Between-Case Standardized Mean Difference

In an attempt the enhance the credibility of SCED research findings to the same level of group-comparison design studies, Hedges et al. (2012, 2013) introduced a statistical model for estimating a design comparable effect size estimate also known as the between-case standardized mean difference effect size for SCEDs (BC-SMD). The BC-SMD is an effect measure that can be interpreted on the same scale as the standardized mean differences from group-comparison designs (i.e., Cohen's d). Researchers are in general familiar with this quantification and therefore interpretations are more straightforward as there is an established scale reflecting what can be considered "small", "medium" and "large". Another advantage of the BC-SMD is that results of SCEDs can be combined with results from group-comparison designs (e.g., Zelinsky & Shadish, 2018), providing more evidence related to the intervention effectiveness investigated through both types of designs, allowing to increase external validity. The use of the BC-SMD in its original version, using moment estimation (Hedges et al., 2012, 2013), is justified when its assumptions and requirements are met: at least three participants with similar data patterns (i.e., the effect is an immediate and sustained change in level in absence of trend), normal distribution of the within-case errors and the between-case variation, and constant within-case variance and auto-correlation parameter across cases. The within-case errors follow a first order auto-regressive term. It should be noted that the BC-SMD can also be estimated using (restricted) maximum likelihood estimation (Pustejovsky et al., 2014; Valentine et al., 2016) with fewer assumptions. Although the original BC-SMD using moment estimation can be conceptualized as

a HLM (Hedges et al., 2013), it is the proposal by Pustejovsky et al. (2014) that uses the same estimation procedure as HLMs.

Hierarchical Linear Models

HLMs are general modeling techniques, which can be thought of as extensions of regression-based techniques (i.e., simple OLS and piecewise regression, Moeyaert, Ugille, et al., 2014). HLM can be conceptualized as an approach resulting in across-case regression-based quantifications. Therefore, the justifications previously provided for the regression-based quantifications are also applicable here. Two-level HLMs can be used when there are several participants included in the same study, as in a multiple-baseline design (Ferron et al., 2009), but also for other SCEDs that involve multiple participants such as a replicated withdrawal/reversal design (Shadish et al., 2013). The unique benefit of using HLM is that it results in the estimate of an overall intervention effect (quantified as change in level and/or change in slope) across participants (expressed in raw or standardized units). This makes it possible to make more generalized inferences about the effect of an intervention. It is necessary to remark that HLM is a modeling technique that can be applied to estimate effect sizes measures. Depending on the HLM specification, different effect size measures of interest can be estimated. Two commonly used HLM parameterizations result in the across participants estimate of the change in outcome level and the change in slope between baseline and intervention conditions (Moeyaert, Ugille, et al., 2014). The underlying estimation procedure is commonly (restricted) restricted maximum likelihood or Bayesian estimation (Moeyaert et al., 2017). The decisions regarding how exactly to model the data in the context of a multilevel model can be made in relation to the measurement characteristics of the outcome (Declercq et al., 2019), according to the specific

SCED used (Moeyaert, Ugille, et al., 2014; Shadish et al., 2013) and according to the linear or nonlinear data pattern expected or observed (Shadish, Zuur, et al., 2014).

In addition to estimating effects, variability in the effectiveness of the intervention between participants can be estimated. This is informative as an intervention might work, in general, but not to the same degree for all the study participants. Multilevel models can also handle data complexities such as autocorrelation, either assuming it similar in the baseline and intervention condition, or treating it as heterogeneous (Moeyaert, Ferron et al., 2014) and count data (Declercq et al., 2019). In terms of performance, this technique has been shown to estimate without bias the overall intervention effects (even with a number of participants as small as three, Ferron et al., 2009), but not the variances (Baek et al., 2020). If the research interest lies in the estimate of the intervention effect across cases, then the two-level HLM is appropriate and recommended, even with a small number of participants. If the research interest lies in capturing the between-participant variability in the intervention effect, then more study participants are needed and even then biases are anticipated.

Regarding the connections between HLMs and other quantitative analysis techniques, it should be noted that the BC-SMD can be estimated using the HLM approach. An extension of the basic model underlying the BC-SMD was proposed by Pustejovsky et al. (2014) for computing a standardized mean difference when trends are modeled and allowed to vary between participants and restricted maximum likelihood estimation is used instead of moments estimation. Moreover, HLM could be used in connection with other quantitative analysis techniques, such as by applying a multi-level meta-analysis model (a particular form of HLM) to within-case effect size quantifications or by using generalized linear mixed models (such as random effects Poisson models), where the effect size metric is a form of log response ratio.

Using Within-Case Quantifications for Aggregation

When several basic effects are evaluated in the same study, it is first necessary to verify what proportion of times the basic effect is replicated, for instance requiring a 3:1 ratio of effects to no effects (Cook et al., 2015). Second, a weighted or an unweighted mean can be used to combine the quantifications obtained for each basic effect (Parker, Vannest, Davis, et al., 2011; Schlosser et al., 2008; Swaminathan, Rogers, Horner, et al., 2014), in order to obtain an aggregate quantification of the intervention effectiveness. Apart from aggregating the effect as an average, it is usually informative to assess the variability of effects within- and between-cases, which can be useful for identifying relevant moderator variables. Another option is to use an across-case quantification, such as the BC-SMD (under the assumption of a similar data pattern across participants) or HLM (with the possibility to quantify the variation across participants). The latter two options are recommended as the inverse of the variance is used as a weight to combine across cases.

The Dimensions

The dimensions presented in the current document are based on the single-case experimental design (SCED) literature, including methodological research (see the section entitled “Overview of Single-Case Quantitative Analysis Techniques” and Appendix A) and published applied research (see Fingerhut et al., 2020). Table 1 includes the dimensions and their facets that can be used for properly justifying the selection of an effect measure as part of the a priori data analytical plan.

Dimension 1: Research Question and/or Type of Quantification Desired

The first dimension refers to the intended use of the data analytical approach. An initial facet to consider is the kind of analysis to perform. Formative analysis is performed as part of the data gathering process itself and is crucial for designs that implement some forms of response-guided experimentation, e.g., deciding when to change the conditions (Connell & Thompson, 1986; Swan et al., 2020). Formative analysis is commonly performed via visual inspection, whereas quantitative techniques are useful for summative analysis post data collection (Ledford et al., 2019). Similarly, it is considered that the assessment of whether a functional relation or experimental control is established is mainly done via visual analysis (Maggin et al., 2018; Wolfe et al., 2016), considering also the close interaction between the researcher and the participants (Perone, 1999).

If the aim is to provide a quantitative summary of the degree of intervention effectiveness, a choice between within-case quantification and an across-case quantification is necessary (Odom et al., 2018; Swaminathan, Rogers, & Horner, 2014). Additionally, for several within-case quantifications (e.g., nonoverlap indices) and across-case quantifications (e.g., BC-SMD and regression-based quantifications) it is possible to focus either on the descriptive information or the inferential information. The descriptive information is the effect measure (i.e., the value of the nonoverlap index, the estimate of the BC-SMD, or the estimates of interventions effects in a HLM), whereas the inferential information is represented by the confidence intervals of these effect measures. Additionally, by using a randomization test, a *p*-value and confidence intervals (Michiels et al., 2017) can be obtained at the within-case level or at the across-case level using the HLM approach. The researcher can decide whether to focus on the descriptive or inferential information on the basis of the aims of the analysis and the tenability of the

assumptions required for the validity of the inferential information (e.g., if independent data is to be assumed or normally distributed residual).

Another facet to take into consideration is the kind of desired summary statistics (e.g., whether to express it in standardized or raw units, Manolov et al., 2014). Intervention effectiveness can be reflected in standardized units (BC-SMD, Hedges et al., 2012, 2013, standardized regression coefficient, Van den Noortgate & Onghena, 2008), as a percentage (e.g., mean baseline reduction, Olive & Smith, 2005; a transformation of the log response ratio, Pustejovsky, 2018), in raw units reflecting the original scale (e.g., unstandardized regression coefficient, Van den Noortgate & Onghena, 2003, or the slope and level change, Solanas et al., 2010). An example in applied literature of an effect measure being used because it is unstandardized or standardized can be found in Good (2019) and Lanovaz et al. (2019), respectively.

Finally, given that it is possible to provide a quantification regarding several data features (i.e., level, trend, variability, immediacy, overlap; Kratochwill et al., 2010, 2013), it is necessary to decide, prior to gathering the data and looking at the most salient data feature, which is the focal data feature (or multiple focal features). An example in applied literature of a quantification being used because it is can measure change in trend can be found in Caron and Dozier (2019).

Dimension 2: Design Features

Several design features are expected to be reported (Tate et al., 2016) and they can be used for informing the selection of an effect measure. On the one hand, certain quantifications are only applicable (or more easily and meaningfully applicable) to certain SCED types. For instance, the BC-SMD has been developed to reflect intervention effectiveness for reversal and

multiple-baseline designs (Hedges et al., 2012, 2013; Pustejovsky et al., 2014), but cannot be applied to alternating treatments designs. The mean absolute distance is a meaningful test statistic only for changing criterion designs (Onghena et al., 2019). HLMs are most easily applicable to multiple-baseline designs (Ferron et al., 2009), although their application to other design structures including a replication across individuals is also possible (Moeyaert, Ugille, et al., 2014; Shadish et al., 2013).

Two other especially relevant design features, replication and randomization, were commented earlier. The kind of replication (within-case or across-case) is related to the type of SCED and also to whether the quantification is performed at the within-case or across-case level. If the researcher expects a considerable variability across cases (due to any differences they might have in terms of the type, severity, or history of the issue treated), it may be less meaningful to summarize the results about different cases in a single effect measure. Regarding randomization, it can enable either a masked visual analysis in the context of response-guided experimentation (Byun et al., 2017; Ferron et al., 2017) or a statistical inference about causality when using a randomization test (Onghena, 2020).

Dimension 3: Data Characteristics

In terms of the expected features of the data, several quantities can be expected to be known a priori, although changes to the initial plan may take place during the course of the study. First, the number of participants is usually pre-established and it is relevant for across-case quantifications, such as the BC-SMD or the HLM approach. For BC-SMD and HLM, a minimum of three study participants is needed and unbiased intervention effects can be expected, but the standard errors are likely to be biased and, thus, *p*-values and confidence intervals are likely inappropriate for such a small number of participants are not appropriate. Thus, the

number of participants is relevant for the precision of estimates and also for statistical power. An example in applied literature of an effect measure being used because of the number of participants can be found in Raulston et al. (2019).

A second relevant quantity is the number of measurements[1] available. On the one hand, the number of measurements per phase can have an impact on the effect measures (Pustejovsky, 2019). On the other hand, this quantity is also relevant for the precision of estimates and also for statistical power (although less important than the higher level units in HLM). In relation to statistical power, for randomization tests it is mainly related to the number of possible randomizations, which is related both to the number of participants and the number of measurements available, but also to the randomization scheme (e.g., see Levin et al., 2018, for multiple-baseline designs and Onghena & Edgington, 2005, for alternating treatments designs). An example in applied literature of an effect measure being used because of the number of measurements can be found in Raulston et al. (2019).

Another facet refers to the measurement characteristics (i.e., the measurement scale) of the outcome. Only nonoverlap indices are applicable to ordinal data and in case the target behavior is measured via a (subjective) rating scale that is only ordinal, effect measures based on means or on trend lines cannot be meaningfully applied. Contrarily, there are certain effect measures such as the log-response ratio (Pustejovsky, 2018, 2019) and the Bayesian response ratio (Natesan Batley et al., 2020) that are applicable only when there is an absolute zero, as when the outcome variable is expressed as a frequency (i.e., a ratio scale). Furthermore, certain

[1] We can distinguish between phase and alternation designs (Onghena & Edgington, 2005). In phase designs, such as multiple-baseline and a reversal, the number of measurements refers to the quantity of data points available in each phase. In contrast, in an alternation design such as an alternating treatments design, the number of measurements refers to the number of data points for a given condition, considering the whole alternation sequence.

effect measures used only for descriptive purposes (e.g., the standardized mean difference and the slope and level change procedure by Solanas et al., 2010) are applicable to both interval and ratio scale measures. However, for standardized mean differences accompanied by standard errors for constructing confidence intervals, it is common to assume that the outcome is measured continuously (Valentine et al., 2016). A distinction between an interval and ratio scale outcome is necessary for modeling techniques such as regression-based quantifications and HLMs, which assume normality and continuous outcome. When the outcome is a count, which is common when direct observation is used for gathering data (Pustejovsky, 2019), modifications in the modeling may be necessary (Declercq et al., 2019). It is also important to consider whether the use of direct observation is accompanied by the more recommendable momentary time sampling (Cook & Snyder, 2020) or by partial interval recording (Pustejovsky & Swan, 2015); partial interval recording leads to a quantification of frequency or to a quantity that is not directly interpretable in terms of either frequency or duration and that may lead to misrepresenting the magnitude of effect for certain effect measures (Ledford et al., 2015; Pustejovsky et al., 2019).

Another facet refers to challenging aspects of the data that can be anticipated, so that a way of dealing with them can also be decided prior to gathering the data. In relation to autocorrelation, certain techniques take it into account (i.e., HLMs, BC-SMD, the GLS approach by Swaminathan, Horner, Rogers, et al., 2014, and the interrupted time series simulation approach by Tarlow & Brossart, 2018), whereas others (e.g., NAP, Tau-U) assume it is absent in order to consider their standard errors valid. Additionally, there are effect measures ignoring autocorrelation and not aiming for any statistical inference (e.g., standard errors, confidence intervals). Examples of such quantifications are mean baseline reduction (Olive & Smith, 2005),

percentage of data points exceeding the median (Ma, 2006), slope and level change (Solanas et al., 2010), and the ratio of distances (Carlin & Costello, 2018).

Regarding the possibility of missing data, several methods are applicable (Hox, 2020: Kwasnicka & Naughton, 2020) and have been tested for SCED data, such as expectation-maximization and multiple imputation (Chen et al., 2020; Peng & Chen, 2018; Smith et al., 2012). A different approach is followed in randomization tests (see De et al., 2020, for a randomized marker approach). In the context of HLMs, to the best of our knowledge a review of how missing data has been handled is available only outside of the SCED context (Dedrick et al., 2009). However, it has been stated that one of the advantages of HLMs is precisely handling missing data (Wiley & Rapp, 2019).

Finally, trend estimation could be compromised by outliers (Vannest et al., 2012) and unequal time intervals between measurement occasions. Although outliers cannot be anticipated with confidence, it may be reasonable to opt for robust effect measures. Additionally, when trend projection takes place (e.g., piecewise regression, GLS, Baseline corrected Tau) an excessively long intervention phase may lead to obtaining impossible predictions (Manolov, 2018; Manolov et al., 2019; Parker, Vannest, Davis, et al., 2011). Finally, unequal time intervals between measurement occasions have been considered an issue in the graphical display of the data, in terms of misrepresenting temporal information if the session number suggests a false uniformity (Kubina et al., 2017). The meaning of time trend is different, according to whether time is represented as session number or, say, calendar days. Unequal time intervals can also have influence on the way in which autocorrelation is to be modelled; specifically, a first-order autoregressive model may not be adequate. Thus, greater caution in the interpretation of trends and autocorrelation is required when data are to be gathered at unequal time intervals.

Dimension 4: Expected Data Pattern

In the context of randomization tests, the test statistic is to be chosen before gathering the data, according to the type of effect expected (Heyvaert & Onghena, 2014a, 2014b; Levin et al., 2017). Here, this idea is extended to other analytical options for SCED data, in line with current recommendations (Maggin et al., 2020). Deciding the analytical plan prior to gathering the data on the basis of the expected data pattern is possible when there is sufficient previous evidence on the specific kind of dependent variable or outcome score and the intervention. Thus, expectations are related to specific outcomes or target behaviors and interventions. For instance, spontaneous improvement prior to introducing an intervention can be expected in rehabilitation (Krasny-Pacini & Evans, 2018; Solomon, 2014, also reports the presence of trend in school interventions) and gradual and slower changes can be expected when measuring academic performance (Maggin et al., 2018). Spontaneous improvement can be represented by an improving baseline trend. It is important to use the information available regarding whether an improving baseline can be expected, as it can be difficult to decide on the basis of the data whether there is a clear trend or not (Chiu & Roberts, 2018). Moreover, in the literature, there are different ways for deciding whether trend should be controlled for (Tarlow & Brossart, 2018): (a) if the trend is stable according to the envelope constructed around it (Lane & Gast, 2014); (b) if the trend estimate is at least 0.20 (Vannest & Ninci, 2015); (c) if the trend is statistically significant (Tarlow, 2017); or (d) always, because trend control is part of the procedure (Solanas et al., 2010). Deciding on the basis of previous knowledge is easier than following a variety of criteria, whose suggestions may not coincide. Similarly, it has to be decided how exactly to estimate trend and how to control for it, given that there are multiple options for both of these steps (Manolov, 2018) and it is not advisable to try out several and selecting the one that is most

favorable (Carlin & Costello, 2018). An example in applied literature of an effect measure being used because it can account for trend in the baseline phase can be seen in Gertler and Tate (2019).

A slower change in the dependent variable or a gradual improvement during the intervention phase is conceptualized as a delayed or progressive effect. Such effects do not need to be discarded, as the latency of the change after the onset of the intervention depends on the type of intervention and domain of functioning (Kazdin, 2019), i.e., the contextual information is crucial (Lieberman et al., 2010). Relatedly, expecting a delayed effect or a transition state between conditions (Brogan et al., 2019) or extinction bursts (Barnard-Brak et al., 2020) can justify focusing on part of the observations obtained; Fisher & Lerman, 2014; Levin et al., 2017; Porcino et al., 2020). Similarly, it is possible to focus on the end of the baseline and the beginning of the intervention phase when an immediate effect is expected (Michiels & Onghena, 2019). Another option is to use models that are specifically applicable to gradual change (Swan & Pustejovsky, 2018; Verboon & Peters, 2020).

Dimension 5: Desirable Features of the Quantitative Analysis Techniques

An effect measure, and a quantitative analysis technique in general, should be statistically sound. This requirement can be operatively defined in different ways, according to the descriptive or inferential use of the effect measure. For description, discriminability between different magnitudes of intervention effectiveness is relevant (Parker et al., 2009; Parker, Vannest, Davis, et al., 2011). Specifically, a problem for discriminability are floor and, mainly, ceiling effects (e.g., the impossibility to distinguish between differently effective interventions once complete nonoverlap is achieved). For inference, the lack of bias and the relative efficiency (and, thus, mean square error) of the estimate are some of the desirable features that are usually

assessed (e.g., Hedges et al., 2012, 2013; Manolov & Solanas, 2013; Moeyaert et al., 2017; Swan & Pustejovsky, 2018), as well as confidence interval coverage (e.g., Baek et al., 2020; Ferron et al., 2009). A different set of inferential statistical properties refers to null hypothesis significance testing. Specifically, Type I error rates (i.e., false positives) and statistical power (i.e., true positives) are commonly assessed (e.g., Borckardt et al., 2008; Declercq et al., 2019; Levin et al., 2018; Michiels & Onghena, 2019). An example in applied literature of an effect measure being used because it does not demonstrate a ceiling effect can be seen in Ginns and Begeny (2019).

Except for randomization tests (Craig & Fisher, 2019), the inferential information requires that the quantification has a known approximate sampling distribution. Specifically, a known sampling distribution makes possible standardizing (Swaminathan, Horner, Rogers, et al., 2014; Van den Noortgate & Onghena, 2008), constructing confidence intervals, and using inverse variance weighting for meta-analysis (Parker, Vannest, Davis, et al., 2011; Shadish et al., 2014). Such knowledge about the sampling distribution comes at the price of certain assumptions about the data or the residuals (Hedges et al., 2012, 2013; Moeyaert, Ferron, et al., 2018; Pustejovsky & Swan, 2018). An example in applied literature of an effect measure being used because it has a known sampling distribution can be seen in Garwood et al. (2019).

Beyond the aforementioned statistical properties, a desirable feature can be defined in terms of its performance relative to other effect measures (e.g., in terms of consistency with visual analysis, correlation with other quantifications, or lack of sensitivity to potentially irrelevant procedural details). Such comparison studies have been performed for nonoverlap indices (e.g., Chen et al., 2016; Wolery et al., 2010; Yucesoy-Ozkan et al., 2020), regression-based quantifications (Brossart et al., 2006) and for several quantifications of different kinds (e.g., Barton et al., 2019; Campbell, 2004; Pustejovsky, 2019). An example in applied literature

of an effect measure being used because it correlates with another quantifications can be seen in Lanovaz et al. (2019).

In relation to the strong tradition of using visual analysis in SCED research (Maggin et al., 2018; Ledford et al., 2019), it is important for the quantifications and any potential transformation of the data to be easily represented visually. This is relevant for effect measures as diverse as nonoverlap indices (Tarlow, 2017) and regression-based quantifications (Declercq et al., 2020; Moeyaert, Ugille, et al., 2014; Parker et al., 2006). Note that the requirement is not necessarily for the quantifications to correlate well with the decisions made by visual analysts, as their performance may not be optimal (Ninci et al., 2015).

Finally, it should be noted that this Dimension 5 may be difficult to apply, as there is not sufficient evidence regarding the performance of all effect measures proposed for SCED data analysis. For instance, to the best of our knowledge, no simulation study has been performed yet on the mean baseline reduction (see Campbell, 2004, for a field test), the ratio of distances (Carlin & Costello, 2018) or on the interrupted time series simulation (Tarlow & Brossart, 2018). Finally, it is not feasible for a single text such as the current one to summarize all the evidence available on all possible effect measures.

The Flowchart

Transforming the Dimensions and Facets into a User-Friendly Flowchart

The selection of the effect measure and the justification of this selection is made easier by using the flowchart (Figure 1) rather than Table 1. However, the flowchart is a simplification of Table 1, as it is based on some, but not all, facets included in the table. The dimensions can be understood as an integration of methodological aspects to be kept in mind, whereas the flowchart

simplifies the set of dimensions and facets and is designed as a decision tree that can readily be used by applied researchers. For instance, a facet of Dimension 2 (Design features) omitted is the presence of randomization. However, we recommend complementing the descriptive information provided by an effect measure with the inferential information provided by a *p*-value arising from a randomization test, when randomization is used. Suggestions for effect measures to be used as test statistics were provided in the "Randomization and Randomization tests" section, previously in the text. Response-guided experimentation (another facet of Dimension 2) is also not discussed, but the interested reader is referred to Byun et al. (2017), Ferron et al. (2017), Joo et al. (2018) and Swan et al. (2020). Finally, certain data characteristics (Dimension 3) such as the level of measurement of the outcome variable, missing data, outliers, autocorrelation are not reiterated here, given that, in general for all effect measures, they entail the need for greater caution in the interpretations. However, for certain quantitative analysis techniques such as HLMs, it is possible to build the model in such a way as to account for count data and autocorrelation.

In summary, when reporting the data analytical decisions made, an applied researcher can report the dimensions and facets that were used as a basis for justifying the quantification chosen (as per Table 1), as well as the pathway followed, according to the flowchart.

INSERT FIGURE 1 ABOUT HERE

The flowchart also illustrates the relations between facets, both within and across dimensions, and also the relation between effect measures. Considering that the focus is put on quantifications for summative analysis, the initial decisions are related to a facet of Dimension 1 (i.e., the unit of analysis), which is necessarily related to two facets of Dimension 2 (the type of SCED and the kind of replication it entails). For ATDs and CCDs there are specific analytical

options and certain potential focal data features (a facet of Dimension 1) such as overlap and or immediacy may not be as meaningful or critical as for other SCEDs.

The expected data pattern is the third crucial aspect in the flowchart, after considering the unit of analysis and the type of design and replication. The expected data pattern (Dimension 4) determines the focal data feature of the quantification (Dimension 1).

Across-Case Quantifications: The Left Pathways in the Flowchart

Given that the presence and type of replication defines the first decision point in the flowchart, a note is required. It is necessary to distinguish across-case replications (i.e., an MBD across participants, a replicated reversal design, or ATD) from within-case replications (i.e., an MBD across behaviors or settings, a reversal design, or an ATD for a single participant). HLMs (including the BC-SMD) are conceptually applicable only to replications across cases. For within-case replication, within-case quantifications can be used.

Following the path for across-case quantifications (to the left of the flowchart), the type of design (a facet from Dimension 2), the anticipation about the similarity across cases (a facet from Dimension 3), and the expected data pattern (Dimension 4) are relevant for assessing whether the BC-SMD is a reasonable quantification. In case the heterogeneity across cases is to be quantified and the presence of baseline or intervention phase trend is considered likely, the BC-SMD could be substituted by a less restrictive HLM such as the one proposed by Pustejovsky et al. (2014). Moreover, HLMs incorporating separate trend lines for the different conditions and random effects are applicable beyond MBDs and replicated reversal designs (e.g., to ATDs and CCDs, Shadish et al., 2013), unlike the BC-SMD. In relation to Dimension 1, the focus on the descriptive information (estimates of treatment effect, as immediate effect and effect

on time trend) or on the inferential information (*p*-values and/or confidence intervals) is a decision to be made by the researcher, but focusing on the descriptive information requires less assumptions. Model building (e.g., the decisions regarding whether to include general trend and the effect of the intervention on the time trend, and which effects to model as random) could be guided by the visual analysis of the actually obtained data (Baek et al., 2016). Another option which we recommend, following Ferron et al. (2008), is to select the model a priori on the basis of the expectations and previous evidence. Nevertheless, it is still possible to plan a post hoc verification that the initially chosen model is meaningful for the data actually obtained and that it does not represent a gross misspecification. In that sense, any subsequent changes in the model need to be explicitly labeled as data-driven. Finally, in relation to Dimension 3 (Data characteristics) and Dimension 5 (Desirable features of the quantitative analysis techniques), it has to be considered whether the application of the BC-SMD or a HLM is reasonable considering the number of participants and the number of measurements finally available.

Within-Case Quantifications: The Right Pathways in the Flowchart

Alternating Treatments and Changing Criterion Designs

When the unit of analysis is the participant, within-case quantifications can be used, following the path to the right of the flowchart. The decision about the quantification depends, first, on the design used. For a changing criterion design, the range-bound version and the percentage of conforming data (McDougall, 2005) can be appropriate as a quantification of the degree to which the data match the pre-established criteria (see also Manolov et al., 2020, as an alternative way for specifying the acceptable range). For ATDs, the mean difference can be used, or a quantification comparing the data paths directly: a comparison that entails actual and linearly interpolated values, abbreviated ALIV (Manolov & Onghena, 2018).

Multiple-Baseline and Reversal Designs: Variable Data

For multiple-baseline and reversal designs, the expected data pattern (a facet from Dimension 4) defines the focal data feature (a facet from Dimension 1). In case the data are expected to be variable and not well represented by a mean/median or trend line, a nonoverlap index can be recommended (Parker, Vannest, & Davis, 2011). Another reason for using a nonoverlap index could be in relation to the measurement characteristics of the outcome. Specifically, nonoverlap indices are applicable to ordinal data (Parker, Vannest, Davis, & et al., 2011, see also Parker and Hagan-Burke, 2007), whereas a mean difference or a comparison of regression slopes requires interval or ratio scale data. Additionally, regression-based quantifications require parametric assumptions such as a normally distributed residual. In relation to the kind of information to use (Dimension 1) and the knowledge on the sampling distribution of the indices (Dimension 5), we recommend focusing on the descriptive measure, as the inferential information (i.e., the standard errors for obtaining *p*-values or constructing confidence intervals) depends on the unlikely assumption of independent data. For choosing among nonoverlap indices, if there is no expectation for an improving baseline trend, the NAP (Parker & Vannest, 2009) can be used. In contrast, if there is such an expectation, we recommend using the Baseline corrected Tau (BCT), forcing trend correction regardless of the statistical significance of baseline trend, because the statistical power of this test is not sufficient for short baselines (Tarlow, 2017). We do not recommend using the BCT without baseline trend correction when no baseline trend is expected, because this would be equivalent to using Tau as proposed by Parker, Vannest, Davis, et al., (2011), but its interpretation is less straightforward than the interpretation of the NAP. BCT represents an improvement over Tau-U (Parker, Vannest, Davis, et al., 2011), as it provides stronger control for baseline trend and it does not

produce out-of-range values. Our recommendation for BCT over Tau-U is also related to the lack of clarity regarding the exact interpretation of Tau-U (Brossart et al., 2018). Nevertheless, BCT is not flawless, as it corrects for linear trend, estimated using the robust Theil-Sen method, and trend extrapolation may lead to impossible projections (Manolov et al., 2019). Unreliable trends and impossible projections are related to a Dimension 3 facet, namely the number of measurements available in each phase, with a short baseline combined with a long intervention phase constituting a major problem. For that reason, we recommend a planned post hoc verification with the actually obtained data. In case such impossible projections are obtained, the correction of baseline trend may be unreasonable and the value of BCT may not be validly interpretable.

Multiple-Baseline and Reversal Designs: Summarizing Data via Mean or Trend Lines

For multiple-baseline and reversal designs, in case stable data and an immediate effect are expected, a comparison of level is the logical option. A subsequent decision refers to the desired measurement units of the quantification (a facet from Dimension 1). The comparison in level can be expressed as a percentage change (Olive & Smith, 2005), which can be obtained from the log-response ratio (Pustejovsky, 2018), or as the percentage of goal obtained (Ferron et al., 2020). Another option is to quantify the difference in level in standard deviations: this is achieved via the within-case standardized mean difference (Busk & Serlin, 1992) or dividing the estimate of the immediate effect in piecewise regression (Center et al., 1985) by the root mean square error (Van den Noortgate & Onghena, 2003). Finally, the difference in level can be expressed in the same measurement units as the outcome, via the slope and level change procedure (Solanas et al., 2010).

For multiple-baseline and reversal designs, in case an improving baseline trend and/or a progressive effect is expected, there is an additional decision to make. One option is to use an overall quantification that takes into account the difference in level and trend jointly: this is achieved via the GLS approach yielding a regression-based quantification (Swaminathan, Rogers, Horner, et al., 2014), by quantifying the average distance between the projected baseline trend line to the fitted intervention phase trend line. This overall difference can be expressed in raw or standardized terms (referring to a facet of Dimension 1). Given that there is trend extrapolation, just like in the BCT (although the trend line is fitted following a different estimation method), we reiterate our caution in relation to impossible projections, especially for certain combinations of phase lengths (Dimension 3). Another option is to obtain separate quantifications of the change in level and change in slope. This can be achieved via piecewise regression (Center et al., 1985) or the slope and level change procedure (Solanas et al., 2010).

Demonstration of the Usability of the Flowchart for A Priori Justification

The use of the flowchart to provide an a priori justification for the SCED selection as part of the protocol will be demonstrated using the protocol by Clanchy et al. (2019), which is one of the protocols located in the literature review by Fingerhut et al. (2020). Given that no data have yet been gathered or made public, the analytical decision cannot be based on (or biased by) the data at hand. In terms of the design, Clanchy et al. (2019) plan to use an MBD across three participants (sample 1), replicated across three more participants (sample 2). The number of sessions for the baseline, given the staggered introduction of the intervention, will be 5, 8, and 11. For the intervention phase, 12 sessions are planned. Participants will be assigned at random to each of the two studies and, afterwards, at random once again to each of the baseline lengths. Unequal spacing between sessions in the intervention phase is planned (more frequent sessions

in the beginning of the intervention and less frequent sessions in the end), but equal spacing of sessions is expected in the baseline phase. In terms of the data characteristics of interest, Clanchy et al. (2019) explicitly mention the need to take autocorrelation into account and the importance of estimating trends and controlling for baseline trend. In terms of the research question, the authors are interested in studying the consistency of effects across participants.

Following the flowchart presented in Figure 1, it is noteworthy that there is replication across participants, which makes possible the use of across-case quantifications (i.e., the left pathways of the flowchart). The design is an MBD across participants for which both the BC-SMD and more complex HLMs are applicable and both take autocorrelation into account. Given that trend is highlighted, as well as the desire to study the consistency across participants, a HLM including trend and quantifying the degree of heterogeneity across cases is required. That is, the model underlying the BC-SMD assuming stable levels and similarity across cases (Hedges et al., 2012, 2013) is not sufficient, whereas the more recent version by Pustejovsky et al. (2014) would be appropriate. For modeling trend, the time variable can be coded in such a way as to represent real time, instead of just session number (Moeyaert, Ugille, et al., 2014), considering the unequal spacing in the intervention phase. Autocorrelation can be modeled as homogeneous or heterogeneous (Moeyaert, Ferron, et al., 2014). Moreover, using random effects can be included to represent and quantify the variability in the effects across participants. Finally, a moderator variable could be included, to code for each of the six participants whether they belong to sample 1 or to sample 2, in order to check whether there are differences between these samples. Given the presence of randomization in the design, a recent proposal for the combined use of the HLM approach and a randomization test for obtaining *p*-values (Michiels et al., 2020) could be used for data analysis.

Discussion

Building on the Existing Literature

Identifying and improving questionable research practices is necessary (Hantula, 2019). Focusing on SCED data analysis, there have already been suggestions on how to achieve such an improvement, for instance via preregistration (Hales et al., 2019; Johnson & Cook, 2019), by choosing a test statistic on the basis of the effect expected (Heyvaert & Onghena, 2014a, 2014b; Levin et al., 2017), and by being explicit regarding whether any hypotheses were established before or after exploring the data visually (Kwasnicka & Naughton, 2020). The underlying reason for the current text is to help applied researcher to avoid capitalizing on chance, which is especially likely when analyzing the data according to the most salient data features or trying multiple analyses and reporting the one that suggests that an intervention effect is present. For that purpose, the current text proposes a set of dimensions and facets to be used when selecting a SCED quantification, as part of the a priori data analytical plan. These dimensions represent a systematic organization and integration of factors previously mentioned by the founders of the several effect measures, in order to take into account the multiple pieces of information that need to be considered when making analytical decisions.

Recommendations for Applied Researchers

General Recommendations

The following recommendations refer to summative analysis and not to formative analysis for response-guided experimentation or to exploratory research in a domain with no previous empirical evidence. We suggest that applied researchers explicitly refer to each of the dimensions and the relevant facets. The information provided for these dimensions and facets

can be used to follow the flowchart from Figure 1, suggesting a pathway that helps to determine a priori the most appropriate quantification and to justify the selection. If the authors plan to use several quantifications and verify the convergence of conclusions (or to perform a sensitivity analysis), this has to be mentioned as well, in order to avoid selective (and biased) reporting of results, which could lead to overestimating intervention effectiveness.

In certain situations, a change in the data analytical plan can take place. For instance, if there are unexpected modifications during data collection stage or if certain features of the data obtained are considered to invalidate the quantifications chosen a priori. In such cases, reporting both the planned and the post-hoc analyses is recommended. Thus, when reporting the result of the planned analysis, it is necessary to alert the reader that the assumptions of the effect measure are not met or that it is likely to misrepresent the data at hand (whichever is applicable). Complementarily, when reporting the result of the effect measure selected a posteriori, it is necessary to highlight that this measure is not the one initially planned and that its selection may entail a form of overfitting. We also recommend performing a sensitivity analysis, comparing the conclusions that would be reached by the planned and the post-hoc analyses, with the confidence in the conclusions being higher when these conclusions coincide and the need for caution when interpreting the effect measures being greater otherwise.

A distinction between a priori expectations and post-hoc analyses might be relevant for preventing false positives. Moreover, reporting that the findings do not match the expected data patterns can be useful for prompting research into the potential reasons for the unexpected results and for improving the interventions or the fidelity with which they are implemented (Tincani & Travers, 2018). Finally, a comprehensive evaluation of the effect of the intervention needs

include social validity indicators (Horner et al., 2005), beyond the visual inspection and the quantifications.

Recommendations in Relation to the Flowchart

The flowchart presented in the current text is based on the convergence of dimensions and is designed to reflect the intended use of the SCED quantifications, as established by their founders. For a good conceptual understanding of the quantification and its main features, we strongly recommend that applied researchers consult and refer to the original sources. For that purpose, apart from the references provided in the current text, a list of selected publications is made available as Appendix A at https://osf.io/t96fc/?view_only=19c595b82d9c4a96aec04eeaf6a4b196. Additionally, we recommend that applied researchers provide details about the software or tool that will be used for obtaining the quantification of choice. As an aid, we provide a list of selected freely available tools, as Appendix B at https://osf.io/t96fc/?view_only=19c595b82d9c4a96aec04eeaf6a4b196. In summary, we consider that prior to conducting the study, the whole analysis program should be included.

Limitations and Future Research

Regarding specific types of SCEDs, the focus of the current text is put on the most common options, specifically the A-B comparisons which are the building block of MBDs and reversal designs, and also on ATDs and CCDs. For combined designs, quantitative proposals are very recent (Moeyaert et al., 2020). In relation to MBDs across participants, it should be noted that the BC-SMD is applicable for combining intervention effects, but the logic of MBDs usually

requires comparing within-series and between-series (Hayes, 1981). For this latter purpose, the interested reader can consult Ferron et al. (2014) and Joo and Ferron (2019).

The current text refers mainly to summative quantifications in present of previous evidence, but is not intended to suggest that formative analysis or response-guided experimentation is not useful or to state that exploratory research is inappropriate and very specific expectations are always required. Following Simmons et al. (2011), it is suggested that researchers explicitly state when their study is exploratory and, if possible, gather more data, presenting a replication (confirmatory) study with additional participant(s), using the same analytical approach as in the original exploratory study.

In relation to the flowchart, it includes multiple decision points and pathways, because SCEDs can be different and the focus can be put on different data features, which renders the decision making process complex. Nevertheless, the flowchart is still a simplification of the set of dimensions and facets and simplifying it further, in excess, would not represent reality. In terms of the degree of comprehensiveness of the flowchart, we believe that it is reflecting most common pathways, but that researchers using it can probably identify additional pathways. For instance, the current version of the flowchart has omitted the need to deal with nonlinear trends, as many effect measures assume linear trends (Solomon et al., 2015). However, it is not reasonable to expect that all trends are linear and continue unabated in time (Parker, Vannest, Daivs, et al., 2011). In consequence, several analytical options have been discussed for dealing with nonlinear trends: for instance, in the context of generalized least squares regression analysis proposal by Swaminathan, Rogers, Horner, et al. (2014), when using generalized additive models (Shadish, Zuur, et al., 2014), when using multilevel models (Hembry et al., 2014) and in the context of randomization tests (Solmi et al., 2014). A review of these modeling options is beyond

the scope of the current text, but it is necessary to state that any previous evidence on a possible nonlinear trend has be taken into account when deciding a priori how to analyze the data. In absence of evidence for nonlinearity, parsimony would call for initially opting for modeling trend as linear. In case the actually obtained data suggest nonlinear trends, it is important how such nonlinearity can be interpreted from a substantive perspective, before deciding how to analyze the data. For instance, a nonlinearity stemming from a delayed effect (see Levin et al., 2017) needs to be distinguished from a nonlinearity stemming from an effect that reaches an asymptote (see Swan & Pustejovsky, 2018). If an alternative effect measure or modeling technique is decided after the data are obtained, a distinction between the planned analysis (involving linear trend) and a post-hoc analysis (including nonlinear models) would be necessary. This distinction should follow the same rules for reporting as stated in the previous section.

In terms of future research, the current version of the flowchart can serve as an initial step for a dialectic research process in which it is continuously improved by researchers. Any subsequent modifications in the flowchart, on the basis of the experience and expertise of other researchers with different backgrounds (applied, methodological, statistical) is likely to increase its usefulness.

In order to explore the effect of using an effect measures selected a priori versus choosing the quantification to the features of the actually obtained data, a field test would be useful. A comparison of a priori vs. post-hoc analyses can inform about the degree to which there is difference between the two approaches and in what direction. Specifically, it can be assessed whether the planned analyses lead to more conservative quantifications of effect.

References

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

Baek, E. K., Petit-Bois, M., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). Using visual analysis to evaluate and refine multilevel models of single-case studies. *The Journal of Special Education, 50*(1), 18–26. https://doi.org/10.1177/0022466914565367

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

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

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

Beckers, L. W., Stal, R. A., Smeets, R. J., Onghena, P., & Bastiaenen, C. H. (2020). Single-case design studies in children with cerebral palsy: A scoping review. *Developmental Neurorehabilitation*, *23*(2), 73–105. https://doi.org/10.1080/17518423.2019.1645226

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63*(2), 77-95. https://doi.org/10.1037/0003-066X.63.2.77

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., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau-U for single-case experimental designs. *Cogent Psychology, 5*(1), article 1518687. https://doi.org/10.1080/23311908.2018.1518687

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*(5), 531–563. https://doi.org/10.1177/0145445503261167

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.

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

Byun, T. M., Hitchcock, E. R., & Ferron, J. (2017). Masked visual analysis: Minimizing Type I error in visually guided single-case design for communication disorders. *Journal of Speech,*

Language, and Hearing Research, *60*(6), 1455–1466. https://doi.org/10.1044/2017_JSLHR-S-16-0344

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*(2), 234-246. https://doi.org/10.1177/0145445503259264

Carlin, M. T., & Costello, M. S. (2018). Development of a distance-based effect size metric for single-case research: Ratio of distances. *Behavior Therapy, 49*(6), 981–994. https://doi.org/10.1016/j.beth.2018.02.005

Caron, E., & Dozier, M. (2019). Effects of fidelity-focused consultation on clinicians' implementation: an exploratory multiple baseline design. *Administration and Policy in Mental Health and Mental Health Services Research, 46*(4), 445-457. https://doi.org/10.1007/s10488-019-00924-3

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data what they tell us and what they don't. *Behavior Modification, 37*(3), 378–390. https://doi.org/10.1177/0145445513476609

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., Feng, Y., Wu, P.-J., & Peng, C.-Y. J. (2020). Dealing with missing data by EM in single-case studies. *Behavior Research Methods*, *52*(1), 131–150. https://doi.org/10.3758/s13428-019-01210-8

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

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

Chiu, M. M., & Roberts, C. A. (2018). Improved analyses of single cases: Dynamic multilevel analysis. *Developmental Neurorehabilitation, 21*(4), 253–265. https://doi.org/10.3109/17518423.2015.1119904

Clanchy, K. M., Tweedy, S. M., Tate, R. L., Sterling, M., Day, M. A., Nikles, J., & Ritchie, C. (2019). Evaluation of a novel intervention to improve physical activity for adults with whiplash associated disorders: Protocol for a multiple-baseline, single case experimental study. *Contemporary Clinical Trials Communications, 16*, 100455. https://doi.org/10.1016/j.conctc.2019.100455

Connell, P. J., & Thompson, C. K. (1986). Flexibility of single-subject experimental designs. Part III: Using flexibility to design or modify experiments. *Journal of Speech and Hearing Disorders, 51*(3), 214–225. https://doi.org/10.1044/jshd.5103.214

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

Cook, K. B. & Snyder, S. M. (2020). Minimizing and reporting momentary time-sampling measurement error in single-case research. *Behavior Analysis in Practice, 13(*1), 247–252. https://doi.org/10.1007/s40617-018-00325-2

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

De, T. K., Michiels, B., Tanious, R., Onghena, P. (2020). Handling missing data in randomization tests for single-case experiments: A simulation study. *Behavior Research Methods, 52*(3), 1355–1370. https://doi.org/10.3758/s13428-019-01320-3

De Young, K. P., & Bottera, A. R. (2018). A summary of reporting guidelines and evaluation domains for using single-case experimental designs and recommendations for the study of eating disorders. *International Journal of Eating Disorders, 51*(7), 617–628. https://doi.org/10.1002/eat.22887

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

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods, 51*(6), 2477–2497. https://doi.org/10.3758/s13428-018-1091-y

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69-102. https://doi.org/10.3102/0034654308325581

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

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

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

Ferron, J. M., Joo, S.-H., & Levin, J. R. (2017). A Monte Carlo evaluation of masked visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis, 50*(4), 701–716. https://doi.org/10.1002/jaba.410

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. http://dx.doi.org/10.1037/a0037038

Ferron, J., Rohrer, L. L., & Levin, J. R. (2019). Randomization procedures for changing criterion designs. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519847627

Ferron, J. M., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*(2), 167–178. https://doi.org/10.1080/00220973.1995.9943820

Fingerhut, J., Moeyaert, M., & Manolov, R. (2020). *Literature review of single-case quantitative analysis techniques*. https://psyarxiv.com/7yt4g

Fisher, W. W., & Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: Lies, damn lies, and statistics". *Journal of School Psychology, 52*(2), 243–248. https://doi.org/10.1016/j.jsp.2014.01.001

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

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

Garwood, J. D., Werts, M. G., Mason, L. H., Harris, B., Austin, M. B., Ciullo, S., Magner, K., Koppenhaver, D. A., & Shin, M. (2019). Improving persuasive science writing for secondary students with emotional and behavioral disorders educated in residential treatment facilities. *Behavioral Disorders, 44*(4), 227-240. https://doi.org/10.1177/0198742918809341

Gertler, P., & Tate, R. L. (2019). Behavioural activation therapy to improve participation in adults with depression following brain injury: A single-case experimental design study. *Neuropsychological Rehabilitation*. Advance online publication. https://doi.org/10.1080/09602011.2019.1696212

Ginns, D.S. & Begeny, J.C. (2019). Effects of performance feedback on treatment integrity of a class-wide level system for secondary students with emotional disturbance. *Behavioral Disorders, 44*(3), 175-189. https://doi.org/10.1177/0198742918795884

Gonzales, J. E., & Cunningham, C. A. (2015). The promise of pre-registration in psychological research. *Psychological Science Agenda, 29*(8). Retrieved from https://www.apa.org/science/about/psa/2015/08/pre-registration

Good, K. E. (2019). *The pen or the cursor: A single-subject comparison of a paper-based graphic organizer and a computer-based graphic organizer to support the persuasive writing of students with emotional and behavioral disorders or mild autism* (Publication No. 13864282.) [Doctoral dissertation, George Mason University]. ProQuest Dissertations.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science, 42*(1), 13–31. https://doi.org/10.1007/s40614-018-00186-8

Hantula, D. A. (2019). Editorial: Replication and reliability in behavior science and behavior analysis: A call for a conversation. *Perspectives on Behavior Science, 42*(1), 1–11. https://doi.org/10.1007/s40614-019-00194-2

Hayes, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49*(2), 193-211. https://doi.org/10.1037/0022-006X.49.2.193

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

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple-baseline design data. *The Journal of Experimental Education, 83*(4), 514-546. https://doi.org/10.1080/00220973.2014.907231

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

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Hox, J. (2020). Important yet unheeded: Some small sample issues that are often overlooked. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 254–265). Routledge.

Institute of Education Sciences. (2020). *Standards for excellence in education research*. Retrieved from https://ies.ed.gov/seer/index.asp

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

Johnson, A. H., & Cook, B. G. (2019). Preregistration in single-case design research. *Exceptional Children, 86*(1), 95-112. https://doi.org/10.1177/0014402919868529

Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology, 1*(1-3), 1–7. https://doi.org/10.1080/23743603.2015.1070611

Joo, S. H., & Ferron, J. M. (2019). Application of the within-and between-series estimators to non-normal multiple-baseline data: Maximum likelihood and Bayesian approaches. *Multivariate Behavioral Research, 54*(5), 666-689. https://doi.org/10.1080/00273171.2018.1564877

Joo, S. H., Ferron, J. M., Beretvas, S. N., Moeyaert, M., & Van den Noortgate, W. (2018). The impact of response-guided baseline phase extensions on treatment effect estimates. *Research in Developmental Disabilities, 79*, 77-87. https://doi.org/10.1016/j.ridd.2017.12.018

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

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

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

Kratochwill, T. R., 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., & 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., & 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

Krypotos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. *Journal of Abnormal Psychology, 128*(6), 517–527. https://doi.org/10.1037/abn0000424

Kubina, R. M., Kostewicz, D. E., Brennan, K. M., & King, S. A. (2017). A critical review of line graphs in behavior analytic journals. *Educational Psychology Review, 29*(3), 583-598. https://doi.org/10.1007/s10648-015-9339-x

Kwasnicka, D., & Naughton, F. (2020). N-of-1 methods: A practical guide to exploring trajectories of behaviour change and designing precision behaviour change interventions. *Psychology of Sport and Exercise, 47*, 101570. https://doi.org/10.1016/j.psychsport.2019.101570

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

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

Ledford, J. R., Ayres, K. M., Lane, J. D., & Lam, M. F. (2015). Identifying issues and concerns with the use of interval-based systems in single case research using a pilot simulation

study. *The Journal of Special Education, 49*(2), 104–117. https://doi.org/10.1177/0022466915568975

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

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. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation, 21*(5), 290–311. https://doi.org/10.1080/17518423.2016.1197708

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2020). Investigation of single-case multiple-baseline randomization tests of trend and variability. *Educational Psychology Review*. Advance online publication. https://doi.org/10.1007/s10648-020-09549-7

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB…AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50*(5), 599-624. https://doi.org/10.1016/j.jsp.2012.05.001

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. http://dx.doi.org/10.1037/a0018600

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

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

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*(5), 598–617. https://doi.org/10.1177/0145445504272974

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*(2), 287–311. https://doi.org/10.1007/s10864-013-9187-z

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

Maggin, D. M., Robertson, R. E., & Cook, B. G. (2020). Introduction to the special series on results-blind peer review: An experimental analysis on editorial recommendations and manuscript evaluations. *Behavioral Disorders, 45*(4), 195–206. https://doi.org/10.1177/0198742920936619

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. (2019). A simulation study on two analytical techniques for alternating treatments designs. *Behavior Modification, 43*(4), 544–563. https://doi.org/10.1177/0145445518777875

Manolov, R., Gast, D. L., Perdices, M., & Evans, J. J. (2014). Single-case experimental designs: Reflections on conduct and analysis. *Neuropsychological Rehabilitation, 24*(3-4), 634–660. https://doi.org/10.1080/09602011.2014.903199

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

Manolov, R., & Moeyaert, M. (2017b). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

Manolov, R., & Onghena, P. (2018). Analyzing data from single-case alternating treatments designs. *Psychological Methods, 23*(3), 480–504. https://doi.org/10.1037/met0000133

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology, 51*(2), 201–215. https://doi.org/10.1016/j.jsp.2012.12.005

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

McDougall, D. (2005). The range-bound changing criterion design. *Behavioral Interventions, 20*(2), 129–137. https://doi.org/10.1002/bin.189

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49*(1), 363–381. https://doi.org/10.3758/s13428-016-0714-4

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

Michiels, B., Tanious, R., De, T. K., & Onghena, P. (2020). A randomization test wrapper for synthesizing single-case experiments using multilevel models: A Monte Carlo simulation study. *Behavior Research Methods, 52*(2), 654–666. https://doi.org/10.3758/s13428-019-01266-6

Moeyaert, M., Akhmedjanova, D., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2020). Effect size estimation for combined single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 14*(1-2), 28-51. https://doi.org/10.1080/17489539.2020.1747146

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

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

Moeyaert, M., Zimmerman, K. N., & Ledford, J. R. (2018). Synthesis and meta-analysis of single-case research. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology. Applications in special education and behavioral sciences* (3rd ed.) (pp. 393-416). Routledge.

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

Natesan Batley, P., Shukla Mehta, S., & Hitchcock, J. H. (2020). A Bayesian rate ratio effect size to quantify intervention effects for count data in single case experimental research. *Behavioral Disorders*. Advance online publication. https://doi.org/10.1177/0198742920930704

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

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification, 39*(4), 510–541. https://doi.org/10.1177/0145445515581327

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nuzzo, R. (2015). How scientists fool themselves–and how they can stop. *Nature News, 526*(7572), 182. https://doi.org/10.1038/526182a

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. (2020). One by one: The design and analysis of replicated randomized single-case experiments. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 87–101). Routledge.

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

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

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., & Hagan-Burke, S. (2007). Single case research results as clinical outcomes. *Journal of School Psychology, 45*(6), 637–653. https://doi.org/10.1016/j.jsp.2007.07.004

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

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*(2), 135−150. https://doi.org/10.1177/001440290907500201

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., & 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

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

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*(2), 109-116. https://doi.org/10.1007/BF03391988

Porcino, A. J., Shamseer, L., Chan, A. W., Kravitz, R. L., Orkin, A., Punja, S., Ravaud, P., Schmid, C. H., & Vohra, S. (2020). SPIRIT extension and elaboration for N-of-1 trials: SPENT 2019 checklist. *BMJ, 368*, m122. https://doi.org/10.1136/bmj.m122

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

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods, 24*(2), 217−235. https://doi.org/10.1037/met0000179

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368−393. https://doi.org/10.3102/1076998614547577

Pustejovsky, J. E., & Swan, D. M. (2015). Four methods for analyzing partial interval recording data, with application to single-case research. *Multivariate Behavioral Research, 50*(3), 365−380. https://doi.org/10.1080/00273171.2015.1014879

Pustejovsky, J. E., & Swan, D. M. (2018). *Effect size definitions and mathematical details*. https://cran.r-project.org/web/packages/SingleCaseES/vignettes/Effect-size-definitions.html

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

Raulston, T. J., Zemantic, P. K, Machalicek, W., Hieneman, M., Kurtz-Nelson, E., Barton, H., Hansen, S. G., & Frantz, R. J. (2019). Effects of a brief mindfulness-infused behavioral parent training for mothers of children with autism spectrum disorder. *Journal of Contextual Behavioral Science, 13*, 42-51. https://doi.org/10.1016/j.jcbs.2019.05.001

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.

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2*(3), 163-187. https://doi.org/10.1080/17489530802505412

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

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

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology, 52*(2), 149–178. https://doi.org/10.1016/j.jsp.2013.11.004

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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

Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy, 43*(3), 679–685. https://doi.org/10.1016/j.beth.2011.10.001

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34*(3), 195–218. https://doi.org/10.1177/0145445510363306

Solmi, F., Onghena, P., Salmaso, L., & Bulté, I. (2014). A permutation solution to test for treatment effects in alternation design single-case experiments. *Communications in Statistics - Simulation and Computation, 43*(5), 1094-1111. https://doi.org/10.1080/03610918.2012.725295

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*(4), 477–496. https://doi.org/10.1177/0145445513510931

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education, 24*(4), 438–458. https://doi.org/10.1007/s10864-015-9221-4

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*(2), 213–230. https://doi.org/10.1016/j.jsp.2013.12.002

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

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

Tate, R. L., Rosenkoetter, U., Wakim, D., Sigmundsdottir, L., Doubleday, J., Togher, L., McDonald, S., & Perdices, M. (2015). *The risk-of-bias in N-of-1 trials (RoBiNT) scale: An expanded manual for the critical appraisal of single-case reports*. Sydney, Australia: Author.

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

Valentine, J. C., Tanner-Smith, E. E., & Pustejovsky, J. E. (2016). *Between-case standardized mean difference effect sizes for single-case designs: A primer and tutorial using the scdhlm web application*. Oslo, Norway: The Campbell Collaboration. https://doi.org/10.4073/cmdp.2016.1

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

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*(3), 142–151. https://doi.org/10.1080/17489530802505362

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development, 93*(4), 403–411. https://doi.org/10.1002/jcad.12038

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

Vannest, K. J., Peltier, C., & Haas, A. (2018). Results reporting in single case experiments and single case meta-analysis. *Research in Developmental Disabilities, 79*, 10–18. https://doi.org/10.1016/j.ridd.2018.04.029

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

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*(2), 235–268. https://doi.org/10.1353/etc.2012.0010

What Works Clearinghouse. (2020). *What Works Clearinghouse Standards Handbook, Version 4.1*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/handbooks

Wiley, R. W., & Rapp, B. (2019). Statistical analysis in small-n designs: Using linear mixed-effects modeling for evaluating intervention effectiveness. *Aphasiology, 33*(1), 1–30. https://doi.org/10.1080/02687038.2018.1454884

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., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement on the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification, 40*(6), 852–873. https://doi.org/10.1177/0145445516644699

Yucesoy-Ozkan, S., Rakap, S., & Gulboy, E. (2020). Evaluation of treatment effect estimates in single-case experimental research: Comparison of twelve overlap methods and visual analysis. *British Journal of Special Education, 47*(1), 67-87. https://doi.org/10.1111/1467-8578.12294

Zelinsky, N. A. M., & Shadish, W. R. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation, 21*(4), 266–278. https://doi.org/10.3109/17518423.2015.1100690

Table 1

Dimensions and Facets for Justifying the Selection of Effect Measures

DIMENSIONS	FACETS	CHOICES WITHIN FACETS
Research question or type of quantification desired	Formative or summative analysis	Formative analysis: Visual analysis Summative analysis: Quantification
	Presence of functional relation or quantification	Functional relation: Visual analysis Quantification: Effect measure
	Unit of analysis to which the research question refers	Individual analysis: Within-case quantification Aggregation: Across-case quantification
	Descriptive or inferential	Focus on the effect size or on the *p*-value or confidence interval
	Measurement units of the effect measure	Raw (same measurement units as the outcome variable) or comparable (standardized, percentage)
	Focal data feature	Choose one (level, trend, variability, immediacy, overlap) or state explicitly that several features will be quantified, looking for converging
Design features	Type of design	Multiple-baseline design, reversal, alternating treatments design, changing criterion design, combined
	Replication	● Within-case (e.g., reversal design) or across-case (e.g., multiple-baseline design) ● An across-case replication can be inherent to the design (multiple-baseline design across participants) potentially leading to an across-case quantification or additional (reversal, alternating treatments design, changing criterion design, multiple-baseline design across behaviors and across participants) potentially entailing several within-case quantifications

		● Anticipated variability in treatment effectiveness across cases
	Randomization	Present or not; to use in the analysis via a randomization test or not to use
	Response-guided experimentation	Is there a pre-established control for false positives?
Data characteristics	Number of units of analysis	● Depends on whether a within-case quantification or an across-case quantification is to be used. ● Consider whether the number of units is sufficient according to the evidence available for the analytical technique.
	Number of measurements per phase or condition	● Are summaries of level and trend expected to be reliable? ● Are standard errors expected to be estimated with precision? (effect on *p*-values and confidence intervals) ● Will there be enough statistical power?
	Outcome variable(s) scale(s)	Ordinal? Interval (continuous)? Ratio scale (counts)?
	Anticipated challenges (not found post hoc)	● Autocorrelation: Effect measure assumes independence? Does it handle autocorrelation? Evidence on the performance when there is autocorrelation? ● Missing data ● Outliers ● Potential impossible projections of baseline trend ● Unequal time intervals between observations
Expected data pattern	Baseline data pattern	● Expectations about stability (variability) and the usefulness of a summary measure of level ● Need to model time trend
	Intervention effect	Immediate effect vs. progressive or delayed effect

Desirable features of the quantitative analysis techniques	Statistical properties	● Adequate levels of Type I error rates and statistical power; confidence interval coverage; bias and Mean Square Error when estimating ● Better performance than another quantification ● Discriminability when applied to real data
	Known sampling distribution (under certain assumptions)	● Standardizing ● Constructing confidence intervals ● Possibility for inverse variance weighting relevant for quantitative integrations
	Quantifications easily represented visually	● Main quantifications and summaries are easily represented on the time series plot? ● Data transformations or trend corrections are easily represented on the time series plot?

Figure 1

Flowchart for Selecting an Effect Measure According to Several Dimensions