Quantifying differences between conditions in single-case designs:

Possible analysis and meta-analysis

Running head: Quantifying differences in single-case designs

Abstract:

The current paper is a call for and illustration of a way of closing the gap between basic research and professional practice in the field of neurorehabilitation. Methodologically, single-case experimental designs and the guidelines created regarding their conduct are highlighted. Statistically, we review two data analytical options: (a) indices quantifying the difference between pairs of conditions in the same metric as the target behavior and (b) a formal statistical procedure offering a standardized overall quantification. The paper provides guidance in the analysis and suggests free software in order to illustrate, in the context of data from behavioral interventions with children with developmental disorders, that informative analyses are feasible. We also show how the results of individual studies can be made eligible for meta-analyses, which are useful for establishing the evidence basis of interventions. Nevertheless, we also point at decisions that need to be made during the process of data analysis.

Key words: single-case designs, effect size, meta-analysis, standardized mean difference

Research and professional practice need to keep in touch if the former is to be really useful for the latter and if professionals want their everyday work to contribute to constructing scientific knowledge. We consider that there are two gaps that need to be bridged between research and practice: (a) in terms of how data are gathered and (b) in terms of how data are analyzed. In the following, we first briefly discuss some guidelines for collecting data in a rigorous way by means of single-case experimental designs (SCED). Afterwards we focus on the main topic of the current paper – the use and interpretation standardized and raw average difference indices. These indices are presented in the context of other options for data analysis, whose mains strengths and requirements are also mentioned. In order to discuss the indices in more detail, within-study analysis and across-studies meta-analysis are carried out and commented. The analyses presented illustrate of some challenges faced by applied researchers and indications are offered about how to begin coping with these challenges.

Closing the Gap between Research and Practice: Gathering Data

One of the ways in which an intervention can be implemented and its effect assessed via a methodologically rigorous procedure is through SCEDs. Such designs entail recording a behavior of interest repeatedly, before and after an intervention is introduced. SCEDs have already been suggested (Graham, Karmarkar, & Ottenbacher, 2012) and actually used (Perdices & Tate, 2010) in research in rehabilitation. In that respect, McMillan (2013) mentions one aspect that may boost the use of SCED in this domain – the possibility of tailor made interventions –, as well as one condition for including single-case studies as a tool for identifying effective interventions – their good quality. In terms of applicability, SCED scan be considered well-suited for translating practice into research. Actually, the AB design is similar to the natural process of an initial assessment followed by a change in the conditions and continued measurement of the

same behavior of interest (Rabin, 1981). Moreover, SCEDs allow studying low prevalence problems, disperse populations, and focusing on individual clients, with the possibility to modify the intervention according to the client's responses (Edgington, 1983) or even to terminate any harmful interventions (Johnston & Pennypacker, 2009).

Nevertheless, the research designs require more than single switch from evaluation to intervention, given that AB designs are not considered sufficient for demonstrating intervention effectiveness (Tate et al., 2013). Such designs do not offer guarantees for ruling out alternative explanations for behavioral change (e.g., history), which is one of the three main criteria for causality, together with the need for the cause to precede and covary with the effect (Shadish, Cook, & Campbell, 2002). For dealing with this issue, at least three attempts to assess whether a change in the conditions is associated with a change in the target behavior are required. This requirement can be met, for instance, using multiple-baseline designs (MBD), which replicate the AB sequence across different participants, behaviors or settings (sometimes generally referred to as "tiers"), with the intervention for each AB comparison being introduced at a different moments in time. Other recommended, but less frequently used (Shadish & Sullivan, 2011) design structures include ABAB (within-case replication of the introduction and withdrawal of the intervention) and alternating treatment designs with a faster and more frequent change in the conditions (Kratockwill et al., 2010). Apart from choosing an appropriate design structure, choosing the moments of change in phase at random can further help ruling out alternative explanations and boost scientific credibility (Kratochwill & Levin, 2010). In order to make its implementation feasible, the random assignment of the intervention start point can be made after the baseline data have stabilized (Heyvaert & Onghena, 2014).

In order to meet McMillan's (2013) requirement for "good quality" single-case research (p. 793), specific proposals have already been made for assessing quality. For instance, Horner and colleagues (2005) have proposed a set of quality indicators for the special education field, including requirements for a detailed description of participants and settings, precise definition of dependent and independent variables, as well as indicating how internal, external, and social validity can be improved. With a similar aim, Reichow, Volkmar, and Cicchetti (2008) developed a set of quality indicators for evidence based practice in autism. Fortunately, the criteria converge with the ones present in Horner et al. (2005). One of the main differences is that procedural fidelity (Ledford & Gast, 2014) is separated from the definition of the independent variable (as in the latter case it would be restricted to "treatment fidelity"). Another distinction is that visual inspection is added as a desired means of analysis (with statistical analysis present as a quality indicator for group design studies). A review applying the indicators proposed by Reichow et al. (2008) and Kratochwill et al. (2010) for assessing the quality of behavioral interventions in autism spectrum disorder (Camargo et al., 2014) suggested that most of the criteria were met by most of the studies included, although aspects such as design strength and treatment fidelity were only met without reservations by approximately half of the studies, pointing at aspects that still need improvement.

Several specific proposals have been put forward for assessing the quality of a SCED (Smith, 2012). One of them is the methodological quality RoBiNT scale (Tate et al., 2013), useful for assessing research that has already been conducted and reported, and also for guiding the decision-making process while the research is still on-going. On the other hand, the soon-to-be-available guidelines from the SCRIBE project (Tate et al.,

2014) can improve not only the reporting of SCED studies, but also how these studies are carried out, thanks to the specific aspects that these guidelines focus on.

Closing the Gap between Research and Practice: Analyzing the Data

The second gap that needs to be bridged refers to the distance between the analytical proposals made over the last decade and the actual SCED data analysis practices. Despite the evidence from the beginning of the century that visual analysis is still the most frequently used analytical method (Parker & Brossart, 2003), there is already evidence from the neurorehabilitation domain suggesting that statistical analysis is also being frequently used (Perdices & Tate, 2010). Nevertheless, the variety of current developments summarized in several Special Issues (e.g., Barker, Mellalieu, McCarthy, Jones, & Moran, 2013; Burns, 2012; Evans, Gast, Perdices, & Manolov, 2014; Shadish, 2014) still need to make their way into the public space, so that applied researchers learn about the existence, use, and interpretation of these analytical techniques.

Regarding data analysis, the WhatWorks Clearinghouse standards (Kratochwill et al., 2010) stress the usefulness of metrics such as proportions or rates (when available), as well as the convenience of regression-based estimators and a comparison between those and nonparametric indices. Using the Standards as a basis, the RoBiNT scale (Tate et al., 2013) highlights as appropriate either systematic visual analysis, or visual analysis complemented with quasi-statistical techniques, or the justified use of statistical procedures. Accordingly, we here illustrate the joint use of visual analysis and two descriptive or quasi-statistical procedures[1]– mean phase difference (MPD; Manolov &

[1] We use the term "quasi-statistical" here given that it is employed in the methodological quality scale developed by Tate and colleagues (2013). This term refers to procedures that offer quantitative summaries of the data (i.e., a statistical description), but lack the possibility to obtain inferential results (e.g., confidence intervals) on the basis of standard errors which quantify the uncertainty in the point estimate. Thus, such quasi-statistical procedures do not meet the requirement of reporting confidence intervals about the effect size measures (Wilkinson & The Task Force on Statistical Inference, 1999),

Solanas, 2013, with the modification from Manolov & Rochat, 2015) and slope and level change (SLC; Solanas, Manolov, & Onghena, 2010). We also use a proper statistical procedure – the *d*-statistic (Hedges, Pustejovsky, & Shadish, 2012, 2013).

Our choice of analyses is based on the idea that these analytical techniques are likely to be correctly used and interpreted by applied researchers and due to the fact that they offer complementary information: (a) in terms of the metrics: in the measurement units of the target behavior (MPD and SLC) and in standardized terms (*d*-statistic); (b) in terms of the data aspects taken into account: the *d*-statistic deals with autocorrelation and can express the results when trend is not controlled for (if the professional is not sure about the presence or the stability of trend), whereas the other two techniques automatically control for linear trend, but not for autocorrelation; and (c) in terms of the object of the quantifications: MPD and SLC provide separate values for each AB-comparison, whereas the *d*-statistic yields an overall quantification across the replications present in the study. Nevertheless, the reader should be alerted that these are not the only possibilities for SCED data analysis and several other options will be commented in the Discussion.

Closing the analytical gap is possible if practitioners become familiar with the alternative methods (and when each is most useful), but it also requires software implementations that make their application sufficiently easy. Following previous developments in SPSS (Shadish & Marso, 2013), SAS (Moeyaert, Ferron, Beretvas, & Van Den Noortgate, 2014), or R (Brossart, Vannest, Davis, & Patiences, 2014; Bulté & Onghena, 2012; Shadish, Hedges, & Pustejovsky, 2014), we refer (in the Appendix) to code in R created for performing the analyses presented here.

---

although the correctness of the standard errors and confidence intervals of inferential statistical techniques is subjected to the completion of their underlying assumptions.

In the following sections, we illustrate, in the context of published neurorehabilitation data, the information that the techniques chosen provide, as well as the challenges they present. We hope that the illustration and accompanying discussion help convince applied researchers that it is possible to carry out sound analysis (according to the criteria by Tate et al., 2013) and make the results of a single study eligible for meta-analysis that can help building the evidence basis of interventions. We also hope to encourage researchers to pay close attention to all aspects of the data.

Method

Data Selection

Given that the aim of the current article was to illustrate recent analytical developments to a real neurorehabilitation study, any such study would be appropriate, as the analytical techniques need to be applicable to all kinds of situations, if we are to consider them useful. Therefore, in September 2014, we just carried out a hand search of the recent articles of *Developmental Neurorehabilitation*, looking for a study using a SCED, without further restrictions. The most recent study we identified was one carried out by Ninci and colleagues (2013) and it aimed to improve, via a behavioral intervention, the eye contact between a therapist and a 4-year-old boy called Felix with pervasive developmental disorder-not otherwise specified. This study was conducted following a multiple-baseline design, which appears to be illustrative of the higher relative frequency of these design structures (Shadish & Sullivan, 2011; Smith, 2012). The study was presented by its authors as a replication of a previous research (Foxx, 1977), which is why the latter was also selected. Foxx's (1977) study is described as a continuation and extension of a previous study by Foxx and Azrin (1973), also using overcorrection as a behavioral intervention, and this is the reason for selecting this

study. Specifically we here focus on study 2 reported by Foxx and Azrin (1973) aiming to reduce different kinds of self-stimulation in relatively long time periods. We decided to include this third study as well in our meta-analytical integration on the basis of the similarity in the type of intervention and participant characteristics (not on the basis of the target behavior, which is different) in order to illustrate (a) how to proceed when some studies aim to reduce the behavior of interest and others to increase it, and (b) that both the quasi-statistical quantifications and the statistical procedure used here can handle replicated ABAB data, but that they do it differently. Finally, note that the three studies do not constitute a random or a representative sample of the research in the domain of behavioral interventions for children with developmental disorders; we rather chose somewhat related studies that illustrate the possibilities and challenges of meta-analysis.

In the Ninci et al. (2013) study there are three tiers (i.e., three replications of the AB structure), one for each of three therapists, with independent and prompted eye contact as target behaviors. For all three replications the study takes place in a childhood playroom and toys are used as reinforcers, which makes the study potentially more ecologically valid. Given the staggered introduction of the treatment (the intervention starts after 4 baseline measurements occasions for Therapist 1, 7 for Therapist 2, and 9 for Therapist 3; with n_B being equal to 12, 9 , and 6, respectively), Ninci et al.'s (2013) study uses a MBD.

In the Foxx (1977) study there are also three tiers, one per participant (called Mike, Wilma, and Doug, who are 8, 8, and 6 years old, respectively), with two therapists also being present and active in the setting. Foxx (1977) describes the design as simultaneous treatment combined with changing criterion. The main interest is the effect of a functional movement training (i.e., an overcorrection). The condition with

overcorrection also includes edibles and praise as reinforcers, whereas the comparison condition only includes the latter two aspects. For the current analysis, we will consider this comparison condition as a baseline, although it does include treatment, but not the treatment of interest (overcorrection). Therefore, for Therapist A, the phase lengths are as follows $n_A = 4$ and $n_B = 17$ for Mike, $n_A = 23$ and $n_B = 6$ for Wilma, and $n_A = 4$ and $n_B = 20$ for Doug; for Therapist B - $n_A = 21$ and $n_B = 7$ for Mike, $n_A = 7$ and $n_B = 22$ for Wilma, and no intervention (and no comparison possible) for Doug. More information regarding the data selected for the analyses is presented below.

Data Analysis

Within-study analysis. According to our view on how SCED data should be treated, it is necessary to always take into consideration three aspects: substantive or clinical significance, the graphical representation of the data, and the numerical summaries that can be obtained from them. For deciding on practical significance, we consider that practitioners are best-suited to assess the presence and magnitude of improvement in the client, according their knowledge of this person and his/her situation, as well as according to their professional experience.

Regarding visual analysis, the majority of evidence (e.g., Danov & Symons, 2008; Ottenbacher, 1990; Ximenes, Manolov, Solanas, & Quera, 2009; but see Kahng et al., 2010 for an exception) indicates that visual analysts may not agree frequently enough. Therefore, complementing naked-eye analysis with visual aids[2] is a reasonable practice (Fisher, Kelley, & Lomas, 2003). Moreover, visual aids can be considered part of the process of systematic visual analyses currently required by the existing standards

[2] See Bulté and Onghena (2012) for a discussion of visual aids such as lines presenting phase means or medians, trend lines fitted to each phase, range lines representing the amount of data variability and also for software tools.

(Kratochwill et al., 2010) and methodological quality assessment tools (Tate et al., 2013). Specifically, it is important to initially focus on the baseline and assess whether it presents certain stability or any improving or deteriorating trend. As a visual aid helpful in this process, the split-middle trend can be mentioned (Miller, 1985). The fitted split-middle trend would indicate, although less precisely than other options such as running medians (Tukey, 1977), whether the baseline is stable or not. Furthermore, one of the steps of systematic visual analysis requires comparing the projection of the baseline data with the actually obtained measurements during the following intervention phase, as when using the MPD. Finally, it has been suggested that data variability around the baseline trend (regardless of whether it is present or flat) needs to be considered via a stability envelope (Gast & Spriggs, 2010). We propose using 1.5 times the baseline phase interquartile range as a measure of variability for constructing this envelope, an option closely related to exploratory data analysis (Tukey, 1977). In absence of trend, the stability envelope or the tools of statistical process control (Callahan & Barisa, 2005) may be appropriate.

Regarding the quantitative analysis, we propose and illustrate a combination of raw indices, expressed in the same measurement units as the data themselves, and standardized indices, as a classical statistical approach. Raw indices can be considered useful for helping applied researchers decide on the practical relevance of the change (i.e., in relation to the clinical significance mentioned above) as they are more directly interpretable, whereas standardized indices favor the comparison and integration of results from outcomes on different metrics in the same study or in different studies (Cumming, 2012).

MPD and SLC both estimate linear baseline trend as the average increase from one measurement to the next one. (In case there is no linear baseline trend, this is reflected

in the estimate and no correction is performed.) MPD projects the estimated baseline trend into the intervention phase data and compares it to the actually obtained data, a comparison suggested as part of systematic visual analysis. Thus, the raw mean difference that MPD quantifies is between projected and actual intervention phase measurements. Regarding SLC, the linear baseline trend estimated is removed from the data, being subtracted from all the (baseline and intervention phase) measurements according to their position in the order in the data series. Afterwards, using the detrended data, the trend still present in the intervention phase is estimated as the average increase from one measurement to the next one. This latter quantification represents change in slope: the average difference in the increase/decrease rate, per measurement occasion, in the intervention phase as compared to the baseline phase. After removing the intervention phase trend from the intervention data, the net level change is computed as a difference between average of the detrended baseline phase data and the average between the doubly detrended intervention phase data. The joint use of these procedures answers Beretvas and Chung's (2008) call for separate estimation of different effects and Swaminathan, Rogers, and Horner's (2014) emphasis on the need for a quantification of the overall effect. The quantifications of both MPD and SLC can be standardized by dividing them by the standard deviation of the baseline phase measurements (Manolov & Rochat, 2015).

The general idea underlying the *d*-statistic by Hedges and colleagues (2012, 2013) is that the average difference between the baseline and intervention conditions (i.e., the raw measure of change) is divided by an estimate of the data variability. This latter estimate takes into account the within-case and between cases variance, and autocorrelation. Moreover, the standardized mean difference estimate is corrected for small sample bias. The idea is relatively straightforward, but understanding the

computations involved in the several formulae presented in Hedges et al. (2012, 2013) requires more advanced statistical knowledge. Given that both within-case and between-cases variance is taken into account, the index is comparable to standardized mean differences obtained from group-design studies. Moreover, the standard error of the index allows constructing confidence intervals as well as using the inverse variance as weight in meta-analysis. In case trend is deemed to be present in the data, detrending is necessary before using the *d*-statistic, as it is not incorporated automatically in the procedure. The statistical model underlying the index assumes: (a) that the change in level is constant across cases; (b) within-case residuals and between-case variation do not change over time and are normally distributed; (c) within-case errors follow a first-order autoregressive process (Shadish et al., 2014). The normality assumption can be tested using the relatively more powerful Shapiro-Wilk test (Razali & Wah, 2011), but Shadish et al. (2014) highlight that unbiased estimates of effect are obtained even in absence of normality. The *d*-statistic index requires at least three cases. In that sense, although it is common for SCED studies to include more than one participant (Shadish & Sullivan, 2011), this requirement still excludes part of the studies (for instance, the ones that use a single participant following an ABAB design and allowing for a within-case replication to demonstrate the experimental effect; Kratochwill et al., 2010).

Across-studies analysis. When performing a meta-analysis, it is necessary to deal with any possible dependence in the outcomes, especially when several outcomes are obtained from the same study), because it is assumed that the outcomes combined are independent. (See Cheung & Chan, 2004, for a general overview and a proposal for taking dependence into account). One option is to avoid such dependences by using a single effect size per study (Lipsey & Wilson, 2001): averaging of the effect sizes in a study or picking one of those at random or due to a substantive reason (Borenstein,

Hedges, Higgins, & Rothstein, 2009). Another option is to model the dependence. With the advent and adaptation of multilevel models to SCED data it is possible to use all outcomes obtained within a study and to take into account the nested structure of the data (effects within studies) and the dependencies that arise from it (Van den Noortgate & Onghena, 2003, 2008). Multilevel models would thus avoid the need for making (sometimes) arbitrary decisions about how to obtain a single effect per study and has also been shown to yield appropriate standard errors and interval estimates of the effects, even without the need to know in advance the amount of dependence, as assumed by multivariate meta-analytical models (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). The option followed here, as we are not using multilevel models, is to obtain a single quantification of effect per study.

The *d*-statistic yields directly a single quantification for a MBD or a replicated $(AB)^k$ design (including replicated ABAB), and thus for a study. In contrast, MPD and SLC were initially proposed for comparing only a pair of phases, as in an AB design. This difference illustrates a distinction between some analytical techniques that handle complex design structures more directly (see also Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014, for multilevel models and Levin, Ferron, & Kratochwill, 2012, for randomization tests) and other analytical techniques such as nonoverlap indices for which several proposals have been made regarding their application to design structures more complex than AB: Ross and Begeny, 2014, compare techniques only in MBD data sets for which there is a single AB for each tier; Parker et al., 2011, use only the initial AB comparison from all design structures; and Olive & Smith, 2005, recommend comparing the initial baseline to the final intervention condition.

In order to obtain a single MPD or SLC effect size per study, the weighted average of the quantifications for each AB-comparison is computed, using the number of

observations within a comparison as a weight. Focusing on the AB-comparisons is consistent with Scruggs and Mastropieri's (1998) recommendation to perform only comparisons that maintain the A-B sequence and with Parker and Vannest's (2012) caution regarding a possible incomplete return to baseline levels in the withdrawal phase of an ABAB design, pointing at the possibility to omit the B_1A_2 comparison for the calculation. At the within-study level, the weight for comparison j is computed as $w_j = n_A + n_B$ and the effect size is $ES_{study} = \sum_{j=1}^{comparisons} ES_j \cdot w_j / \sum_{j=1}^{comparisons} w_j$. Despite the fact that these weights do not capture the influence of all possible nuisance parameters (e.g., autocorrelation, intraclass correlation, Hedges et al., 2013), their use has been suggested by Shadish, Rindskopf, and Hedges (2008) and Kratochwill et al. (2010), when the variance of the estimator is unknown.

In order to obtain a single effect size per study, our decisions are explained here. Regarding the Foxx and Azrin (1973) study, the *d*-statistic can be computed for three replicated ABAB designs (Barbara, Wilma, and Tricia), but not for Mike for whom only AB data are available – another option would have been to use the initial AB for all four participants. Thus, for MPD and SLC we also omitted the data for Mike to ensure comparability and obtained the weighted average of all six[3] AB comparisons (two per participants). The Foxx and Azrin (1973) data are primarily included for the meta-analytical purpose without paying in-depth attention to data patterns, as the Study 2 data suggest clear effects (high baseline self-stimulation reduced to 0 during the intervention phase) even when inspected only visually. Regarding the Foxx (1977) study, the

[3] Note that neither the MPD nor the SLC, in this application, take into account the fact that the six AB comparisons belong to three (rather than six) participants. Such nesting is taken into account by the *d*-statistic and would also be taken into account by a multilevel model. The same results for MPD and SLC we obtained would have been obtained via the following steps: (1) obtain a weighted average per participant, with the weight being the number of measurements per AB comparison; (2) obtain a weighted average per study out of the effects per participant, with the weight being the number of measurements per participant; (3) obtain the weighted average across studies, using series length and the inverse of the coefficient of variation, as explained.

recommendation for using the *d*-statistic when there are at least three replications of AB sequences in a MBD made us focus on the data for Therapist A and exclude the data for Therapist B. We focus on the same data for the MPD and SLC to make the results comparable, obtaining the weighted average for the three participants. Regarding the Ninci et al. (2013) study, we focus only on prompted eye contact, as it is the type of behaviour studied in Foxx (1977) and it is also the data that are more interesting for our illustrative purposes as they allow pointing at situations in which the visual aids should be used with caution. Nevertheless, independent eye contact is also relevant for the substantive purpose of the study, as it is likely to be the ultimate goal for this target behavior (i.e., that the child becomes autonomous), although eye contact can be considered a mere prerequisite for teaching other more complex behaviors such as speech. Finally, a reasonable doubt can be raised regarding whether the results of the studies by Foxx and Azrin (1973) and Foxx (1977) are completely independent, given that they share one participant (Wilma, as Mike's data is from 1973 is not included in the meta-analysis).

Once a single effect size per study is obtained, it is also important to assign a weight to this effect size according to the amount of information available. The *d*-statistic, being based on a solid statistical theory, allows using the inverse of the index variance as a weight, as is common in the meta-analysis of between-group studies. For MPD and SLC this option is not available and the weight was proposed to be a function of the amount of measurements available in the study and the inverse of the variability of the outcomes (Manolov & Rochat, 2015). Specifically, for each study k, $w_k = \sum_{i=1}^{comparisons}(n_{Ai} + n_{Bi}) + \frac{1}{CV'_k}$, where $CV'_k = \frac{\sqrt{\sum_{j=1}^{tcomparisons}(ES_j - ES_{study})^2/tiers}}{|ES_{study}|}$. The idea of incorporating the within-study variability of effects is related to Hershberger,

Wallace, Green, and Marquis' (1999) proposal for using what they call "replication effect" quantification as a moderator. It has to be noted that this weight has not been derived analytically and, thus, it is not as statistically solid as an inverse variance weight. For those researchers that consider that it not necessary or justified to include the information about the variability of effects and who consider that the impact of $\frac{1}{CV_k'}$ is likely to be too small to be relevant[4], Manolov and Rochat (2015) proposed and illustrated using $n_{Ai} + n_{Bi}$only as a weight.

Finally, note that we have referred to MPD and SLC so far as raw indices expressed in the same metric as the target behavior. The three studies review here all use percentages (trials with eye contact in Ninci et al., 2013, and Foxx, 1977, and time samples with self-stimulation in Foxx and Azrin, 1973) and it is thus possible to integrate their results without any transformation. However, given that it is not likely that all studies included in a meta-analyses use the same measurement units, we will also illustrate how to apply the standardized versions of MPD and SLC; for a percentage-version see Manolov and Rochat (2015).

Results

Analysis of Individual Studies

Visual analysis. Figure 1 contains the Ninci et al. (2013) data for prompted eye contact with added visual aids in the form a split-middle trend estimated from and fitted to the baseline and projected into the intervention phase. Recall that this projection is made as an interval of values, defined according to the variability (1.5 times the

[4] A preliminary study that we carried out showed that another possible weight $\sum_{i=1}^{comparisons}(n_{Ai} + n_{Bi}) \times \frac{1}{CV_k'}$ giving greater importance to $\frac{1}{CV_k'}$, actually lead to too large differences between the weights assigned to different AB comparisons.

interquartile range) of the baseline data. The improvement of the target behavior is evident, as no treatment phase measurements enter into the interval of values expected in case the intervention was ineffective. These data illustrate that the visual aids are to be interpreted with common sense. First, for Therapist 1, no data are included in the trend stability envelope, but in the end of the series there is actually a deterioration as compared to what is expected if the baseline trend progressed unchanged and, thus, measurements out of the interval predicted is not necessarily equivalent to improvement. Second, for Therapist 3, the projection includes impossible negative values and, therefore, appears to be an inappropriate reference.

INSERT FIGURE 1 ABOUT HERE

For the Foxx (1977) data as gathered by Therapist A (Figure 2), the intervention phase behaviors are also increased with respect to what is expected in case baseline trends are maintained. The visual aids show that even for Mike and Wilma, for whom there is greater variability in the baseline phase and, therefore, less certainty in the exact values expected, the measurements obtained are out of the intervals that would suggest no change in the behavior. However, the short and variable baseline phase for Mike presents the challenge of deciding whether trend can be estimated with sufficient precision from such data. In summary, the behavioral change here seems clearer than for Ninci et al. (2013), given that the effect is sustained (rather than temporary) for all three replications.

INSERT FIGURE 2 ABOUT HERE

Quasi-statistical and statistical analyses. Both Foxx (1977) and Ninci et al. (2013) focus on average levels per condition and ranges of values as the only quantifications on which they base their assessment of intervention effectiveness, although Foxx (1977)

does mention a change in the behavior with time when commenting on the data sets. However, we will see that they use other indicators of effectiveness not related to statistics. Therefore, taken together the numerical and substantive evidence, it can be argued that the analyses performed by Foxx and Ninci et al. are sufficient for their (within-study) purpose of demonstrating the effectiveness of the behavioral intervention. Still, it has to be stressed that the data analysis method used in both studies is very similar, despite more than 30 years of distance and despite the existence of new and promising analytical techniques. We focus on such techniques here as they build on the basic information of the difference in means and make possible computing effect size indices, which are necessary for quantitative integrations useful for establishing the evidence basis of interventions (Jenson, Clark, Kircher, & Kristjansson, 2007).

As far as the numerical analysis is concerned, the raw MPD and SLC values are presented in Table 1, whereas their standardized versions can be found in Table 2.

INSERT TABLES 1 AND 2 ABOUT HERE

For the Ninci et al. (2013) data on prompted eye contact, MPD suggests that, for all three replications the actual intervention phase measurements are, on average, higher than the ones expected in case baseline trend continued. For Therapist 1, the average difference is only 7% which agrees with the visual representation, suggesting a crossing between projected baseline trend and actual intervention phase trend. For the remaining two therapists the average difference between phases is around 30%, which also agrees with visual impression of clearer effect. SLC provides more detailed information in its two estimates, with the slope change estimate suggesting for all three replications that the intervention phase trend is deteriorating with respect to the baseline trend. This was clearly noted for Therapist 1. For Therapist 2 the baseline is flat, whereas the

intervention phase shows, on average, a decreasing trend: according to the estimate it decreases, on average, with 7.88% per measurement occasion. For Therapist 3, the baseline trend is decreasing, but in the intervention phase this decrease is even more pronounced. According to the results of the quantifications of change in slope it would appear that the intervention led to worse effects. Nevertheless, the net change in level, after eliminating all linear trends, is highly positive: more than 50% average increase between conditions. This information quantifies the visual impression that there is a clear change in level, but that the effect of the intervention is not progressive (does not continue improving with time) or even maintain at the same level (as the negative estimates for the slope change suggest).

The *d*-statistic summarizes the information about all three replications in a raw mean difference equal to 37.51% and a standardized mean difference corrected for small sample bias equal to 2.71 (standard error ≈ 0.60). Other pieces of information used by the *d*-statistic and provided as output are the autocorrelation estimate of 0.33 (incidentally, very similar to the average autocorrelation for multiple baseline designs studies reported in the review by Shadish & Sullivan, 2011: 0.32, indicating that the data are not independent) and an intraclass correlation equal to 0 (that is, all the variation in observations is within-therapists not between-therapists). On the one hand, the raw mean difference (37.51%) is greater than the weighted average for MPD (24.57%), probably related to the fact that MPD controls for trend, relevant for the data for Therapist 1. On the other hand, the standardized value of the *d*-statistic (2.71) is smaller than the weighted average of the standardized MPD values (3.75), probably related to the fact that the latter is standardized according to the (relatively smaller) baseline variability, whereas the former takes into account the variability in all observations.

For the Therapist A data from the Foxx (1977) study, the MPD yields very high quantifications: for two of the participants the increase in prompted eye contact is greater than 100%. This result can be explained by taking into account the fact that baseline trend is not fitted via the split-middle method (as shown on Figure 1), but as the average increase or decrease of successive measurements. This method leads to a negative trend being estimated for all three cases (see Figure 3) and, if projected, this trend "predicts" negative percentages for the intervention phase. We have included this graph and these results to alert applied researchers using procedures controlling for trend, as short and variable baselines like Mike's may lead to such opposed estimates of trend. Taking into account that trend is estimated in SLC in the same way as in MPD, it is not surprising that in all cases a positive change in slope is found (approximately 4% increase per measurement occasion during the intervention for the three participants). Moreover, there is a large net change in level, which is clearer for Wilma (87.66%) for whom the baseline phase measurements are lower than for the other two participants.

INSERT FIGURE 3 ABOUT HERE

The *d*-statistic summarizes the information about all three replications in a raw mean difference equal to 63.25% and a standardized mean difference corrected for small sample bias equal to 4.17 (standard error ≈ 0.86). Other pieces of information used by the *d*-statistic and provided as output are the autocorrelation estimate of 0.45 (once again suggesting that autocorrelation should be taken into account) and an intraclass correlation equal to 0. In this case, the raw value of the *d*-statistic (63.25%) is smaller than the weighted average for MPD (94.74) and the standardized value (4.17) is also smaller (MPD=12.05). The results for MPD are influenced by: (a) the projection of the baseline trends into very low (or even impossibly negative) intervention phase values, which are then compared to the actual high intervention measurements; and (b) the low

variability in the baseline for Doug (i.e., a very small the denominator), which contributes to having a very large average standardized difference.

Regarding the Foxx and Azrin (1973) data, due to space limitations, we will not go into detail reviewing the results presented in Tables 1 and 2. We only mention that the raw *d*-statistic is equal to −67.63%, the standardized one to −4.42 (standard error ≈ 1.01), autocorrelation = 0.56 and intraclass correlation = 0.32 suggesting certain variation across cases. Note that given that the aim of the study was to reduce the target behavior (self-stimulation) and the intervention was effective, practically all quantifications have negative signs. These signs had to be reversed prior to carrying out the meta-analytical integration of results, so that a positive outcome always means the treatment was effective in improving the outcome

Assessment of practical significance. Obtaining evidence on the clinical significance of any behavioral change is a crucial part of data analysis. In the current section we review the indicators of clinical significance used by Foxx (1977) and Ninci and colleagues (2013) and make some suggestions for additional assessment. First, the design used by Foxx already helps ensuring practical significance as the criterion for "adequate performance" changes according to the improvements observed in the participant (i.e., a glance is required initially and at least a 2-second eye contact in the end). The same role has the criterion established by Ninci et al. (2013), requiring that the prompts are provided consistently until there is eye contact in at least 80% of the opportunities, plus the fact that the intervention phase for Therapist 3 was not terminated until the participant reached 70% independent eye contact.

Second, another planned aspect of the studies was the generalization training in Foxx (1977) and the maintenance measures obtained by Ninci et al. (2013), one and

three months post-intervention. In the Foxx study, the generalization training took place in a more natural setting, such as the day care program and led to all children reaching 90% eye contact and the fading out of the reinforcers (edibles and praise).

Third, Foxx (1977) reports that after several intervention sessions certain behaviors incompatible with attending the teacher (e.g., bouncing on the chairs, pushing the table) stopped occurring. This is another indicator of the effectiveness of the program applied, as incompatible behaviors can be seen both as a tool (when they are reinforced in order to replace problematic conduct), and as a nuisance, when they stand in the way of the desired target behaviors.

Finally, for the maintenance measures, it could be useful to relate them to normative measures, such as the ones used in the initial assessment of the 4-year-old Felix, which place him in the 0-18 months group according to social interaction skills. It would be interesting to check the age equivalence of his behavior one and three months after the intervention. Another possibility is to evaluate the degree of overlap of the maintenance measure(s) with the baseline data (maintenance performance should be better) and to the intervention phase data (performance should be similar). Such a comparison could be performed using the Nonoverlap of all pairs (Parker & Vannest, 2009) or even the same *d*-statistic. The application of MPD and SLC is less clear here, as they take baseline trend into account and thus the requirement for comparing only adjacent phases (Gast & Spriggs, 2010) seems crucial.

Quantitative Integration of Several Studies

As explained before, MPD and SLC quantify AB-comparisons, which afterwards need to be averaged in order to obtain a single effect size per study – the results of this process, using the amount of measurements in each AB-comparison as a weight, is

available in Tables 1 and 2. Once a single effect size per study is available, a weight can be assigned to this effect size. For MPD and SLC, the amount of measurements in the study and the inverse of the variability of the outcomes are used as elements of the weight. If we apply the formula for the weight to the standardized MPD outcomes for the Ninci et al. and the Foxx data, we observe that the relative variation of outcomes is approximately equal (59%) and thus the whole difference in weights is due to the number of measurements available:

$$w_k = \sum_{i=1}^{tiers} (n_{Ai} + n_{Bi}) + 1 \Big/ \frac{\sqrt{\sum_{j=1}^{tiers} \left(ES_j - ES_{study}\right)^2 / tiers}}{\left|ES_{study}\right|}$$

$$w_{Ninci} = (16 + 16 + 15) + 1 \Big/ \frac{\sqrt{(6.30 + 8.29 + 0.15)/3}}{|3.75|} = 47 + \frac{1}{0.591} = 48.69$$

$$w_{Foxx} = (21 + 29 + 24) + 1 \Big/ \frac{\sqrt{(0.74 + 55.20 + 94.67)/3}}{|12.05|} = 74 + \frac{1}{0.588} = 75.70$$

The modified forest plot representing the MPD effect sizes expressed in the original metric (i.e., percentages) can be seen on Figure 4. We refer to this graphical representation as a modified forest plot, given that the intervals for the effects do not represent confidence intervals (the standard error of MPD is not known), but rather the range of the outcomes within the study for study effects and the range of effects across studies for the weighted average. The size of the square boxes still represents the weight of the study effect size, but this weight is not based on the inverse variance, but rather on the formula for w_k presented previously. Finally, we have chosen to order the studies according to their effect sizes in ascending order, which can also be done in a traditional forest plot, if the order is not chronological or alphabetical. From Figure 4 it can be seen that the weighted average difference between the projected baseline trend and the actual

intervention phase measurements (82%) is closer to the results of the studies by Foxx (1977) and Foxx and Azrin (1973): the ones for which the effect is greater and for which the data series are longer. Despite the within-study variability of effects, there is clearly effect of the behavioral intervention for children with developmental disabilities.

INSERT FIGURE 4 ABOUT HERE

The same interpretation can be given to the results of the standardized MPD index presented on Figure 5. In this case, the weighted average (10.48) indicates that the overall difference between the actually obtained intervention data and the prediction made on the basis of the baseline trend is ten times the variability of the baseline data.

INSERT FIGURE 5 ABOUT HERE

For meta-analyzing the effect sizes obtained via the *d*-statistic (see Figure 6) we used a random effects model, because we assume, as it commonly done, that the variability in the effect observed is due to both random error and true variation (e.g, in this case due to the fact that the target behavior in the Foxx and Azrin, 1973, study is different) and given that random effects models allow making inferences to similar studies that vary in several characteristics beyond the exact people participating. The weighted average *d* once again suggests the effectiveness of the interventions tested in the three studies, with the overall difference between the measurements obtained in the conditions with and without behavioral intervention being equal to 3.57 standard deviations (which here take into account the variation in the observations within and between replications). The 95% confidence interval is [2.41, 4.74], indicating (a) the statistical significance (at the .05 level) of the weighted average as the value of 0 is not included in the interval; and (b) the relatively low precision of the estimate due to the small number of studies being integrated. The estimated variance of the true effect sizes

is τ^2 = 0.42, with the proportion of true heterogeneity out of the total variability observed being rather small I^2 = 39.23%, that is, between the 25% and 50% cut-offs for small and medium heterogeneity. (Similar meta-analytical analyses can be obtained for the *d*-statistic; Shadish et al., 2014).

Although both MPD and the *d*-statistic indicate a large effect of the behavioral interventions, there is a difference in the magnitude. Part of this difference can be attributed to the fact that there appear to be deteriorating trends for most of the data sets (even for Foxx and Azrin; figure not included here) and thus the MPD values become larger. In case of improving trends, it is expected MPD to provide lower values than the *d*-statistic (if data are not detrended prior to using the latter). The difference is potentially also due to how standardizing is carried out.

INSERT FIGURE 6 ABOUT HERE

Discussion

In the current paper we argue for closing the gap between methodological and statistical (basic) research and the studies that professionals carry out every day, so that this applied research can contribute to establishing the evidence basis of treatments. We decided to base the analytical options discussed here on the analytical practices already taking place – paying special attention to the visual representation of the data and averages for the conditions being compared. For that purpose we chose to illustrate procedures that can help visual inspection and that quantify average differences. These procedure go beyond the mere comparison of means, as they allow: (a) projecting baseline trend (or level in case data present no trend) and comparing it to the actually obtained intervention phase data (MPD); (b) controlling for trend and quantifying change in slope and change in level separately (SLC), which is especially relevant in

case these two affects are not in the same direction, as for the Ninci et al. (2013) data for Therapist 1; (c) obtaining the difference in comparable standardized terms, taking into account autocorrelation, and constructing confidence intervals on the basis of strong statistical theory (*d*-statistic); (d) carrying out meta-analysis (MPD, SLC, and *d*-statistic, with the latter being equivalent to classical statistical procedures). Finally, we chose these procedures as MPD and SLC offer very specific quantifications for each AB-comparison, whereas the *d*-statistic provides an overall estimate of effect considering several features of the data.

Our choice of procedures can also be related to the characteristics of the data. The data used from all three studies show practically all of them 0% overlap and thus nonoverlap indices (Parker, Vannest, & Davis, 2011) are not especially useful for quantifying the magnitude of the difference between conditions when there is complete nonoverlap. For instance, the otherwise recommended Nonoverlap of all pairs (Parker & Vannest, 2009) would have yielded the value of 100% nonoverlap, without further distinction of the different magnitudes of effect. Figures 1, 2, and 3 also suggest that controlling for trend (e.g., via Tau-U; Parker, Vannest, Davis, & Sauber, 2011) would probably not have made a difference.

Other Options for SCED data analysis

Apart from taking the data features into account and discarding nonoverlap indices, our choice of procedures was also based on the idea of highlighting practical procedures, although these are not necessarily the only ones appropriate. First, for maintaining practicality, we did not focus on regression models (Swaminathan et al., 2014) and multilevel models (Moeyaert, Ferron, et al., 2014), or the proposal of Pustejovsky,

Hedges, and Shadish (2014) for an effect size index based on multilevel models. All these analytical options require that the researcher makes decisions on what aspects of the data are to be modelled (e.g., kinds of effects expected – changes in level or in slope, relevance of the variation in effects across cases, the way to proceed with autocorrelation, potential need for standardizing the data in case different measurement units are used; Van den Noortgate & Onghena, 2008). The use of such models is advised under supervision from an experienced analyst. Nonetheless, the supervision and/or training pays off if one is willing to model different data features according to the characteristics of the data at hand (e.g., presence of trend or variability in the effects across the cases) or according to a more general theoretical or empirical background (e.g., presence of autocorrelation, curvilinear trends). Moreover, as stated previously, multilevel models can handle several outcomes per study.

Second, we also did not focus on simple procedures offering information in comparable units (percentages, not standard deviations) such as the Mean baseline reduction (Campbell, 2004) or the Percentage reduction data (Wendt, 2009) in order to avoid forcing the researcher to decide whether to use all the data or only the last three measurements, respectively, given that such a choice might sometimes be based on which results match better the research hypothesis rather than on an a priori substantive justification.

Third, we could not use a randomization test (Heyvaert & Onghena, 2014) for the current data given that no random assignment of conditions to measurement occasions had taken place when gathering the data and this is a requirement for the validity of the procedure (Edgington, 1980) and is also necessary for the adequate performance of the procedure (Ferron, Foster-Johnson, & Kromrey, 2003). If random assignment had taken place, randomization test could provide information in terms of statistical significance

and offer the researcher to the possibility to choose the effect size index to be used as a test statistic, although software implementations such as the SCDA plug-in for R (Bulté and Onghena, 2012) include only a limited set of mean difference test statistics.

Cautions Necessary when Analyzing Data

Despite our desire to make data analysis easier for the reader, one of the things to be learned from the analyses presented here is that are still decisions to be made. First, one decision is whether to use a procedure that controls for trend (like MPD and SLC) or not (like the *d*-statistic) and, in case trend is to be controlled, what method to use for estimating it – regression analysis, split-middle common in visual analysis, the method used in MPD and SLC, the method used in Tau-U, the trisplit discussed and promoted by Parker, Vannest, and Davis (2014). For the Foxx (1977) data we saw that in some cases different procedures can lead to very different estimates of trend. In case trend is taken into consideration, the researcher has to decide whether its control or projection is reasonable or out of bounds (Parker, Vannest, & Davis, 2011), as for the Ninci et al. (2013) data gathered by Therapist 3. In order to detect such situations it is critical to interpret the quantitative analysis guided by the visual inspection of the data. Even when the data are visually inspected, the analyst cannot focus only on the visual aids, but also on the scale of the ordinate indicate the values predicted by projecting the trend. Thus we recommending an in-depth visual inspection of the graph, given the amount of data features it can inform about (Parker, Cryer, & Byrns, 2006) and in order to assess how well baseline trend is estimated and fitted. The SCDA plug-in for R described in Bulté and Onghena (2012) includes several options for estimating trend that can help finding the one that approximates the data best. Second, the user has to know the data well enough to decide whether an overall quantification (*d*-statistic) is sufficiently

informative or it is necessary to distinguish the changes in slope and in level (SLC); another option is to compute both.

Another lesson illustrated is about the standardized quantifications obtained. We saw that the values are far away from Cohen's benchmark for a large effect (0.8). In this context, it is necessary to stress that Cohen himself proposed the benchmarks tentatively, until further evidence is available. These results also illustrate the generally accepted opinion that Cohen's interpretative benchmarks are not suitable for SCED data (Parker et al., 2005), which has led the US Institute of Education Sciences (2014) to state that one of its priorities is to establish alternative guidelines for these designs.

Limitations and Future Research

The current paper presents certain limitations, apart from the already highlighted fact that not all possible (and promising) SCED analytical techniques were illustrated. First, the study does not offer a formal comparison of the performance of the three techniques, given that a limited set of studies was used for illustrating some challenges that researcher may have to face. Second, the focus put here is on the quantifications and to a lesser extent on visual analysis. The discussion of clinical importance is left to the professionals, who are better equipped to use substantive criteria than we are. Third, quality indicators were not applied, given that the focus of this already extensive paper was analytical. Nevertheless, it could have been interesting to explore whether methodological and reporting improvements have taken place from the initial study to its replication more than 30 years later. In any case, professionals considering the use of SCED are encouraged to get acquainted with the methodological quality indicators (Horner et al., 2005; Kratochwill et al., 2010; Reichow et al., 2008; Tate et al., 2013), as these indicators are also relevant to the field of neuropsychological rehabilitation.

Finally, we urge methodologists and statisticians to explain the developments they have worked on in a way that would make them understandable and attractive to applied researchers. We also advocate for incorporating these developments in easy to use software (such as the one included in the Appendix) accompanied by explanations of the quantifications obtained. We hope that the current paper serves as an example of such effort to bring these developments closer to their intended users and that these users would try to keep their data analytical knowledge up to date.

References

Barker, J. B., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2013). Special issue on single-case research in sport psychology. *Journal of Applied Sport Psychology, 25*, 1-3.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley & Sons.

Brossart, D. F., Vannest, K., Davis, J., & Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation, 24*, 464-491.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology, 8,* 104-114.

Burns, M. K., (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education, 21*, 175-184.

Callahan, C. D., & Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: The single-subject design reconsidered. *Rehabilitation Psychology, 50*, 24-33.

Camargo, S. P. H., Rispoli, M., Ganz, J., Hong, E. R., Davis, H., & Mason, R. (2014). A review of the quality of behaviorally-based intervention research to improve social interaction skills of children with ASD in inclusive settings. *Journal of Autism and Developmental Disorders, 44*, 2096-2116.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89*, 780-791.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, UK: Routledge.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*(6), 828-839.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235-251.

Edgington, E.S. (1983). Response-guided experimentation. *Contemporary Psychology, 28*, 64-65.

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation, 24*, 305-314.

Ferron, J. M., Foster-Johnson, L., & Kromrey, J. D. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71*, 267-288.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387-406.

Foxx, R. M. (1977). Attention training: The use of overcorrection avoidance to increase the eye contact of autistic and retarded children. *Journal of Applied Behavior Analysis, 10*, 489-499.

Foxx, R. M., & Azrin, N. H. (1973). The elimination of autistic self-stimulatory behavior by overcorrection. *Journal of Applied Behavior Analysis, 6*, 1-14.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199-233). London, UK: Routledge.

Graham, J. E., Karmarkar, A. M., & Ottenbacher, K. J. (2012). Small sample research designs for evidence-based rehabilitation: Issues and Methods. *Archives of Physical Medicine and Rehabilitation, 93*, S111-S116.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*, 224-239.

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*, 324-341.

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

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation, 24*, 507-527.

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

Institute of Education Sciences. (2014). *Request for applications: Statistical and research methodology in education*. Retrieved from http://ies.ed.gov/funding/pdf/2015_84305D.pdf

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 43*, 35-45.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single case designs technical documentation*. In What Works Clearinghouse: Procedures and standards handbook (Version 2.0). Available at http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 124-144.

Ledford, J., & Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychological Rehabilitation, 24*, 332-348.

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*, 599-624.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Manolov, R., Gast, D. L., Perdices, M., & Evans, J. J. (2014). Single-case experimental designs: Reflections on conduct and analysis. *Neuropsychological Rehabilitation, 24*, 634-660.

Manolov, R., & Rochat, L. (2015, July 27). Further developments in summarising and meta-analysing single-case data: An illustration with neurobehavioural interventions

in acquired brain injury. *Neuropsychological Rehabilitation*. Advance online publication. doi: 10.1080/09602011.2015.1064452

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior Modification, 38*, 878-913.

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*, 201-215.

McMillan, T. M. (2013). Outcome of rehabilitation for neurobehavioural disorders. *NeuroRehabilitation, 32*, 791-801.

Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development, 63*, 491-494.

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*, 191-211.

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*, 665-704.

Ninci, J., Lang, R., Davenport, K., Lee, A., Garner, J., Moore, M., et al. (2013). An analysis of the generalization and maintenance of eye contact taught during play. *Developmental Neurorehabilitation, 16*, 301-307.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation, 28*, 283-290.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357-367.

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*, 254-265.

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*, 303-322.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *Journal of Special Education, 48*, 79-91.

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*, 284-299.

Perdices, M., & Tate, R. L. (2010). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19,* 904-927.

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,* 368-393.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabin, C. (1981). The single-case design in family therapy evaluation research. *Family Process, 20,* 351-366.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2,* 21-33.

Reichow, B., Volkmar, F., & Cicchetti, D. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38,* 1311-1319.

Ross, S. G., & Begeny, J. C. (2014). Single-case effect size calculation: Comparing regression and non-parametric approaches across previously published reading intervention data sets. *Journal of School Psychology, 52,* 419-431.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22,* 221-242.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

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*, 123-147.

Shadish, W. R., & Marso, D. M. (2013). *SPSS macros for effect sizes and power in single-case (N-of-1) designs.* Paper presented at the annual meeting of the Society for Research Synthesis Methodology, Providence, Rhode Island, USA.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188–196.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971-980.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510-550.

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34*, 195-218.

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*, 213-230.

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The conduct and report of single-case research: Strategies to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation, 24*, 315-331.

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

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

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*, 576-594.

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*, 1-10.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention, 2*, 142-151.

Wendt, O. (2009). *Calculating effect sizes for single-subject experimental designs: An overview and comparison*. Paper presented at The Ninth Annual Campbell Collaboration Colloquium, Oslo, Norway. Retrieved June 29, 2015 from http://www.campbellcollaboration.org/artman2/uploads/1/Wendt_calculating_effect_sizes.pdf

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology, 12*, 823-832.

Appendix: User-Friendly Code

The following freely available resources in the also free R software have been used in the current paper.

Data analysis of an individual study. First, the visual inspection of the data can be helped using the visual aids available in the "SCDA" plug-in for R (Bulté & Onghena, 2012; http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html). Apart from the tools available in SCDA, we have also used the R code from https://www.dropbox.com/s/5z9p5362bwlbj7d/ProjectTrend.R in order to fit split-middle trend and project it into the subsequent intervention phase. Second, mean phase difference (MPD) and slope and level change (SLC) procedures for analysis and meta-analysis have been explained in Manolov and Rochat (2015), with the R code for MPD available at https://www.dropbox.com/s/g3btwdogh30biiv/Within-study_MPD_std.R and for SLC at https://www.dropbox.com/s/74lr9j2keclrec0/Within-study_SLC_std.R. Third, the d-statistic has been implemented in R in the "scdhlm" package, available from James Pustejovsky's web page: http://blogs.edb.utexas.edu/pusto/software/.

Meta-analysis of several studies. First, the meta-analysis via MPD and SLC can be performed using the R code available at https://www.dropbox.com/s/wtboruzughbjg19/Across%20studies.R. Second, the meta-analysis via the *d*-statistic, as presented here, can be carried out using this R code: https://www.dropbox.com/s/41gc9mrrt3jw93u/Across%20studies_d.R. Shadish et al. (2014) offer further R code for performing meta-analyses with this index.

Use of the resources in R. Shadish and colleagues (2014) explain the use of their code, as mentioned above. For the remaining pieces of code mentioned in this Appendix, there is a step-by-step tutorial called "Single-case data analysis: Software

resources for applied researchers" available at https://www.researchgate.net/profile/Rumen_Manolov and https://ub.academia.edu/RumenManolov. This tutorial offers (a) an initial introduction to R and R-Commander; (b) an explanation of the way in which data should be organized in order to apply the analysis; (c) a visually-guided list of actions that are required from the user so that the code can be downloaded and executed; and (d) a short guide on the interpretation of the results obtained, with the corresponding reference to the original articles presenting each analytical technique.

The results obtained here. In order to obtain the results presented in this paper, no further specific code was created or adapted. Therefore, the interested reader can replicate the analysis using the code mentioned above and following the indications of the tutorial. What is specific are the data set used, especially given that they were retrieved from the graphs of the articles by Foxx and Azrin (1973), Foxx (1977), and Ninci et al. (2013). Therefore, we offer an Excel file (available at https://www.dropbox.com/s/ybvdhf4q2u3q73q/FoxxNinci.Data.xlsx?dl=0 and online supplementary material) with all the data used and the different ways of organizing it, according to the procedure used. For the analysis and meta-analysis using MPD, SLC, and the *d*-statistic, we recommend that, in order to replicate the analysis, the reader saves each Excel worksheet separately as a tab-delimited text file and then load this text file when performing the analysis. For obtaining the graphical representation of the data and the fitted split-middle trend and its projection, it is necessary to modify the corresponding R code (https://www.dropbox.com/s/5z9p5362bwlbj7d/ProjectTrend.R) introducing the values from the "Measurements" column after **score <- c(** and the length of the baseline phase after n**_a <- .** There is also a worksheet for obtaining the weights for MPD and SLC for the data analyzed in the current article.