A comparison of mean phase difference and generalized least squares for analyzing single-case data

Abstract

The present study focuses on single-case data analysis and specifically on two procedures for quantifying differences between baseline and treatment measurements The first technique tested is based on generalized least squares regression analysis and is compared to a proposed non-regression technique, which allows obtaining similar information. The comparison is carried out in the context of generated data representing a variety of patterns (i.e., independent measurements, different serial dependence underlying processes, constant or phase-specific autocorrelation and data variability, different types of trend, and slope and level change). The results suggest that the two techniques perform adequately for a wide range of conditions and researchers can use both of them with certain guarantees. The regression-based procedure offers more efficient estimates, whereas the proposed non-regression procedure is more sensitive to intervention effects. Considering current and previous findings, some tentative recommendations are offered to applied researchers in order to help choosing among the plurality of single-case data analysis techniques.

Key words: single-case; regression analysis; differencing; autocorrelation; trend

Single-case designs have been frequently used in clinical and educational settings, given that they allow studying the behavior of a single unit (an individual or a group) longitudinally, when it is not possible or not desired to employ group designs. In that sense, when the unit to study presents specific features or a specific problem requiring intervention, single-case studies are called for (Barlow, Nock, & Hersen, 2009; Greenwald, 1976). An investigation designed as a single-case study allows not only focusing on the needs of the unit of interest and treating its particular problem, but it also allows gathering scientific evidence on whether the intervention has been shown to have an effect on the behavior measured (Horner et al., 2005). The simplest design involves comparing the existing situation (i.e., the baseline phase denoted by A) to the evolution of the individual or group after a psychological intervention (i.e., the treatment phase denoted by B). The AB design is considered quasi-experimental, given that it does not allow ruling out alternative explanations of the behavioral change (Campbell & Stanley, 1966). In fact, demonstrating treatment effectiveness requires at least three attempts to study the relationship between the change in conditions (e.g., from baseline to intervention or vice versa) according to current standards (Kratochwill et al., 2010). Such attempts can be made, for instance, in the context of multiple-baseline designs (replicating the AB sequence across participants, behaviors, or settings), withdrawal designs (e.g., an ABAB design entailing three changes in phase), or alternating treatments designs, if the behavior studied is susceptible to rapid changes as a result of the conditions alternation.

Visual Analysis and Visual Aids

The assessment of the relationship between the intervention and the target behavior is suggested to be based on visual analysis focusing on the following data features: within-phase variability, level, and trend; the immediacy of the effect; the amount of overlap between data pertaining to different phases; the presence of outliers; and the whole data pattern observed as compared to the expected one (Kratochwill et al., 2010). This recommendation is well-aligned with the idea that no statistical analysis currently available can consider all these data features simultaneously (Parker, Cryer, & Byrns, 2006).

Visual analysis has historically been considered appropriate and sufficient for longitudinal behavioral data analysis (Michael, 1974; Skinner, 1938) rendering statistical analysis necessary solely when enough experimental control is not achieved (Sidman, 1960). Complementarily, when only strong effects are sought for no further analysis seems to be required (Parsonson & Baer, 1986). Apart from identifying effective interventions, visual analysis has had a positive influence on scientific reports, given that it usually entails presenting the raw data gathered, which makes possible further analyses of these data, both visually (so that each analyst can reach his/her own decision) and statistically. In fact, the single-case reporting guidelines currently being developed include the need to report raw data as a requirement (Tate, Togher, Perdices, McDonald, & Rosenkoetter, 2012). Furthermore, a scale for rating the methodological quality of single-case studies (Tate et al., 2008), also includes this point.

Despite the strengths of visual analysis and the evidence on its continued use (Kratochwill & Brody, 1978; Parker et al., 2006), calls have been made throughout the decades on the need to complement visual and statistical analyses (DeProspero & Cohen, 1979; Franklin, Gorman, Beasley, & Allison, 1996; Houle, 2009; Johnston & Pennypacker, 1980). Among the reasons for a joint use can be stated: a) the lack of

formal decision rules in visual analysis (Kazdin, 1982), b) the corresponding lack of objective and replicable outcomes (Robey, Schultz, Crawford, & Sinner, 1999), c) the idea that practitioners would have more confidence on treatment effectiveness when visual and statistical analyses coincide (Tryon, 1982), and d) statistical analysis can help making single-case findings more credible for the scientific community (Parker et al., 2006). This joint application can favor also statistical analysis, when visual analysis is used to choose the statistical technique to apply and to validate statistical analysis' results (Parker et al., 2006).

The need to accompany visual analysis with an objective criterion does not necessarily imply using statistical analysis. An alternative can be found in visual aids such as the split-middle method for approximating progression lines (White, 1972). This method, when used jointly with a binomial test, allows estimating the probability of as many data points in the desired direction (e.g., above the trend line for behaviors which are to be increased). However, the binomial test has been found to be excessively liberal (Crosbie, 1987).

To overcome the deficiency in terms of Type I error rates associated with the binomial test, the conservative dual criterion was developed (Fisher, Kelley, & Lomas, 2003). This latter method adds a mean line to the trend line and moves both lines in the conservative (desired by the professional) direction by a quarter of a standard deviation before carrying out the statistical test. Given that the majority of studies indicate insufficient interrater agreement (e.g., Danov & Symons, 2008; Ottenbacher, 1990; except when using well-trained analysts making dichotomous decisions on experimental control; Kahng et al., 2010), it is important that training in structured criteria has been shown to increase the confidence in the decisions (Hagopian et al., 1997; Skiba, Deno,

Marston, & Casey, 1989) and to decrease Type I and Type II error rates (Colón, 2006; Fisher et al., 2003).

Quantitative Analysis

The evidence-based movement in the single-case context has made emphasis on the need for quantitative summaries of the results, especially for making them available for meta-analysis (Jenson, Clark, Kircher, & Kristjansson, 2007). The methodological quality scale (Tate et al., 2008) also includes "statistical analysis" as one of the items of methodological quality. The integration of single-case studies is especially relevant given the usual lack of random sampling of participants and as generalization is based on replication (Sidman, 1960). Among the possibilities for integrating data, there have already been studies using simple (Schlosser, Lee, & Wendt, 2008) or weighted averages of the primary indicators (Schneider, Goldstein, & Parker, 2008). A proposal for weighting has also been made for primary indicators for which the standard error is unknown (Manolov & Solanas, 2012). However, before carrying out meta-analysis it is important to select the primary indicator for summarizing the results, a choice which has found so far no consensus. In the following, we review some of the main proposals for numerical or quantitative analysis of single-case data[1]. Afterwards we will focus on two procedures – a recently proposed regression-based technique and an alternative to it. The main proposals for single-case data analysis can be grouped into three categories: randomization tests, procedures based on regression analysis, and procedures related to the criteria commonly used in visual inspection.

[1] We use the terms "numerical" and "quantitative" analysis, given that not all procedures discussed are strictly speaking statistical, provided that they are not all used for inferential purposes.

Randomization Tests

Randomization tests have been claimed not to require parametric assumptions or the need for random sampling and have been proposed for obtaining statistical significance values directly from the data (Edgington & Onghena, 2007). These techniques require, for instance, choosing at random the moment in which an intervention starts in an AB design. The test statistic is computed for the actual intervention start point and for all other possible intervention start points. The *p* value is obtained locating the test statistic for the actual data bipartition (in an AB design) in the distribution of test statistic values. Randomization tests have been suggested for a variety of single-case designs including ABAB (Onghena, 1992) and alternating treatments designs (Onghena & Edgington, 1994), whereas recent adaptations have focused on more complex situations (Lall & Levin, 2004; Levin, Lall, & Kratochwill, 2011). Open-source software has been made available for some of these tests (Bulté & Onghena, 2009) enhancing their applicability. The requirement for introducing randomization in the design can be seen as a gain in terms of scientific credibility (Kratochwill & Levin, 2010), but it also makes the design less practical. The main drawback of these tests is insufficient power in most cases, that is, the difficulty to detect existing intervention effects (Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Levin et al., 2011). The question or whether Type I errors are always controlled or not is a matter of how the analysis is actually carried out and the specific way in which phases are alternated within the design (Levin, Ferron, and Kratochwill, 2012; Sierra, Solanas, & Quera, 2005).

Regression-Based Procedures

Among the strengths of regression analysis should be mentioned the possibility to model a variety of data patterns, including linear and quadratic trend, level and slope change (see Huitema & McKean, 2000). As discussed by Simonton (1977) and Gorsuch (1983), regression also allows accounting for serial dependence (i.e., the sequential relationship between the measurements taken on a single unit, also referred to as "autocorrelation"). Regression permits, after meeting the parametric assumptions, estimating the statistical significance of, say, a level change or a slope change coefficient. It is also possible to express the results as R^2 values (i.e., proportion of variability in the behavior explained by the change in the conditions), convertible to standardized mean difference, that is, average difference between phases expressed in standard deviation units (Grissom & Kim, 2012) . The classical regression-based procedures proposed for single-case data analysis have focused on controlling baseline trend (i.e., the systematic improvement or deterioration in the behavior prior to the intervention) before quantifying intervention effectiveness. Unfortunately, the evidence suggests that they yield either too low R^2 values (Gorsuch's, 1983, Trend and Differencing analyses) or excessively high R^2 values (White, Rusch, Kazdin, & Hartmann's, 1989, index and Allison and Gorman's, 1993, model), according to studies using real-life (Brossart, Parker, Olson, & Mahadevan, 2006; Parker & Brossart, 2003) and simulated data (Manolov & Solanas, 2008). Therefore, these procedures do not allow distinguishing well enough between data with and without treatment effect. There is also evidence that regression-based procedures correlate only moderately with visual judgments (Brossart et al., 2006; Brossart, Parker, & Castillo, 2011). Thus, despite the potential strengths of regression analysis no previous procedure had been shown to be appropriate for single-case data analysis. The most recent proposal was presented by

Maggin and colleagues (2011) as a generalized least squares effect size (hereinafter, GLS) controlling for serial dependence. As in White el al.'s (1989) proposal there are two regression equations, one for each phase, but here all the predicted treatment data are compared, instead of focusing solely on the last one. Maggin et al. (2011) illustrate the procedure applying it to real data sets and a useful complement would be to explore via simulation whether this proposal performs better than the previously developed regression-based procedures. This is the reason for centering the present study mainly on the GLS.

Quantitative Procedures Related to Visual Criteria

A third type of quantifications focusses on the visual criteria such as the amount of overlap between measurements belonging to the different phases, slope and level change (DeProspero & Cohen, 1979; Knapp, 1983; Ottenbacher, 1990). Their main advantages are the ease of computation and interpretation. However, most of them provide a descriptive quantification rather than the possibility of statistical inference. The first procedure based on data overlap proposed is the Percent of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), which compares the best baseline measurement (highest or lowest, according to whether the aim is to increase or reduce behavior) to all the treatment phase measurements. Several modifications of the PND were proposed in order to avoid relying on a single baseline data point. The proposal that has proven (Manolov, Solanas, Sierra, & Evans, 2011; Parker & Hagan-Burke, 2007) to perform best of these alternatives to the PND is the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009), as it is not affected by serial dependence and also distinguishes between presence and absence of effect. This procedure compares each

baseline datum to each treatment datum to quantify the percentage of nonoverlap (i.e., the degree to which the treatment measurements are improved as compared to the baseline measurements). Finally, a procedure was proposed for quantifying separately slope and level change (SLC; Solanas, Manolov, & Onghena, 2010). This procedure first estimates baseline trend as the average of the differenced data, that is, a new series is created subtracting each measurement from the following one and the mean of this series is computed. The baseline trend thus estimated is removed from the whole series. Afterwards, the treatment phase slope is estimated and controlled for in the same way in order to compute finally the pure level change (i.e., the average difference between the two phases' measurements). There is evidence that this procedure estimates precisely slope and level changes and is practically unaffected by serial dependence (Solanas, Manolov, & Onghena, 2010; Manolov et al., 2011). In the present study we propose a procedure based on the SLC, with the purpose to explore the gains and losses in the bias and efficiency of the estimation when using a technique simpler than the GLS.

Objectives of the Current Study

The aim of the present study is to test the GLS with data generated with known characteristics (i.e., serial dependence, trend, presence/absence and type of effect) in order to obtain evidence on its performance. This evidence will make possible emphasizing its advantages and limitations in relation to a nonregression procedure we propose in this paper. Finally, the findings on these techniques will be related to the results available on other procedures in order to help applied psychologists deciding which numerical technique to use.

The Procedures Compared

Generalized Least Squares Regression Analysis

The regression procedure GLS (Maggin et al., 2011) can be summarized in the following steps:

1. Apply ordinary least squares to the whole data series using the following model with the original dependent variable Y and the time variable T representing the session number as a regressor: variables $y_t = \beta_0 + \beta_1 \cdot T$. Specifying the vector T with integer values from 1 to n makes possible taking into account linear trend, whereas if T is assigned squared integer values from 1^2 to n^2 quadratic trend can be handled by the procedure. Unless stated otherwise, in this paper, we applied the GLS version assuming linear trend.

2. Apply the Cochrane-Orcutt procedure for estimating and controlling for autocorrelation. The twelfth chapter of the Kutner, Nachtsheim, Neter, and Li (2005) book should be consulted for further details, but the reader ought to know that autocorrelation (ρ) is estimated as the slope parameter of the following regression equation using original and lagged residuals (ε) from the regression in step 1: $\varepsilon_t = \rho \cdot \varepsilon_{t-1} + u_t$. The serial dependence thus estimated is then removed from the whole data series using the following expressions $Y_1' = Y_1 (1 - \rho^2)^{1/2}$ for the first value in the series and $Y_t' = Y_t - \rho \cdot Y_{t-1}$ for the remaining ones. Note that the Cochrane-Orcutt method is not the only alternative for estimating autocorrelation (see also McKnight, McKean, & Huitema, 2000; Solanas, Manolov, & Sierra, 2010), and is not necessarily the most appropriate one as acknowledged by the authors of the GLS (Maggin et al., 2011).

3. Apply ordinary least squares using the following model with the transformed dependent (*Y'*) and regressor (*T'*) variables $y_t' = \beta_0 + \beta_1 \cdot T'$. Note that there is one regression equation for the baseline and another one for the treatment phase.

4. Check whether the residuals from the previous regressions are autocorrelated using the Durbin-Watson test. In case they are, transform once again the already transformed *Y* and *T* variables and repeat the whole procedure. Kutner et al. (2005) recommend using only one or two iterations; in case more are necessary to eliminate the lack of independence between the final residuals, they suggest using a different procedure. In the current study a single iteration was used.

5. Project the baseline phase regression equation to obtain the predicted treatment data. Obtain the predicted treatment data also from the treatment phase regression equation.

6. Compare the two predicted sets of data and compute the average difference. Note that this comparison actually entails comparing straight two lines, the ones fit by the regression equations (step 5).

7. Standardize the mean difference using the variability around the regression lines in the denominator. In the current study, this last step was omitted, given that our goal was to compare the mean difference yielded by the GLS with the one provided by the non-regression procedure described below.

The numerical value yielded after step 6 is the average difference between the treatment data predicted from the treatment phase regression equation and from the baseline phase regression equation (using time as predictor in both), after having removed the autocorrelation estimated from the residuals. Maggin et al. (2011) state that in the estimation of autocorrelation as performed by the GLS it is desirable that: a) both

the serial dependence scheme and the degree of autocorrelation be the same for both phases, and b) the series consist of at least 20 measurements. Apart from this assumption, for testing regression coefficients' statistical significance the residual is supposed to be homoscedastic and normally distributed. The current study partially aims to explore how the GLS performs when these conditions are not met.

Mean Phase Difference

In the present study, the non-regression procedure compared to the GLS is one based on the trend estimation step present in the SLC. Given this similarity, we expect that the current proposal will control linear trend and will also be only slightly, if at all, distorted by the presence of autocorrelation, as was the case for the SLC (Solanas, Manolov, & Onghena, 2010). The proposal estimates the mean difference between treatment measurements and treatment data as projected using the baseline trend. It is hereinafter referred to as "mean phase difference" (MPD). In comparison to the SLC, with the MPD a single quantification of the amount of change is obtained, without distinguishing between level and slope change. Specifically, the MPD involves the following steps:

1. The baseline phase data (n_A measurements) are differenced, that is, a new set of $n_A - 1$ values is obtained subtracting measurement 1 from measurement 2, measurement 2 from measurement 3 and so forth. First-order differencing (i.e., differencing the series only once) has been proposed for making a series stationary, that is, to remove linear trend. In the context of the MPD, differencing is used a step previous to estimating rather than eliminating linear trend. Specifically, the mean of the differenced $n_A - 1$ values is used for fitting a straight line to the baseline data (i.e., estimating slope, the same aim as in regression analysis) and for predicting (step 2)

how the series would continue into the treatment phase in case no behavioral change takes place contingently with the change in phase.

2. The baseline trend is projected into the treatment phase using the expression $y_{n_A+i} = y_1 + \hat{t} \cdot (n_A + i - 1)$. This formula is interpreted in the following way: each treatment measurement (whose position in the whole series is $n_A + 1$, $n_A + 2$, …, $n_A + n_B$) is equal to the initial baseline value (y_1) plus a prediction of what the measurement would be if the estimated baseline linear trend ($\hat{t}$) is continued to measurement's actual position. As it can be seen, to mark this position (i.e., the order of the measurement in the series), $n_A + i - 1$ is used instead of $n_A + i$, due to the fact that the first baseline datum is used to set the initial level.

3. The projected baseline trend is compared to the actually obtained treatment data and the average difference is computed. Note that, in contrast with the GLS, only one of the data sets compared is a straight line (i.e., the projected baseline trend), whereas the other set are the actual treatment data, which would seldom form a perfectly straight line.

Although the procedure is fairly simple and can be calculated by hand, we offer an R (R Development Core Team, 2012) code in the Appendix, which can be useful when longer data series are available. The code also offers a graphical representation of the original and the projected intervention phase data superimposed. The application of the MPD is restricted to those situations in which there is either no trend present (i.e., the data are stable) or there is linear increase or decrease in the target behavior. In both cases, a straight line is estimated from the baseline and is projected into the treatment phase. It is possible to adapt the MPD to consider quadratic trend (e.g., second-order

differencing can be used), but we wanted to keep the procedure simple as it is compared to a more complex one, which enables dealing with quadratic trend.

The rationale of the MPD is similar to the one of the split-middle method (White, 1972). However, the way in which the baseline trend is estimated is different (i.e., through differencing instead of on the basis of medians). The numerical value yielded is the average difference between the actual treatment data and the treatment data predicted from the baseline linear trend estimated through differencing. This interpretation should be compared to the one applicable to the GLS, in which there is not one but two sets of predicted values and the analysis is based on transformed data rather than on the original ones. For the purpose of the present study both procedures are expressed as a mean difference instead of being standardized. Unstandardized measures allow expressing the primary measures in meaningful units (Cumming, 2012; Grissom & Kim, 2012), as will be shown in the next section, but unstandardized measures are not as useful for comparing results among studies, given possible difference in metric or series length which are not taken into account. For the GLS standardizing is already part of the procedure and for the MPD it is also possible to standardize the mean difference between projected and actual data. This can be done dividing the mean difference by the variability around the estimated trend line in the context of the MPD, which is similar to the solution mentioned in Maggin et al. (2011) for the GLS.

In the following the two procedures are first applied to real data sets and afterwards to generated data in order to obtain evidence on their usefulness and performance. The next section focuses on real psychological data, whereas in the Method section the characteristics of the simulated data are explained, before presenting the Results from the simulation study.

Application to Real Psychological Data

In order to illustrate the two procedures and how they complement visual analysis, in the following section they will be applied to two published psychological data sets. However, a comparison with only two data sets is clearly insufficient for drawing any meaningful conclusions. The Monte Carlo study described later implying simulating data is intended to complement the application to real-life data in order to make possible a more extensive comparison.

The first data set (Figure 1) we will use for the illustration was gathered and reported by Drager and colleagues (2006). The participant is a child referred to as Maggie, who is diagnosed with autism. The behavior of interest is symbol production measured in three different contexts both before and after applying aided language modeling at a different moment for each context.

INSERT FIGURE 1 ABOUT HERE

Inspecting visually all three baselines it appears that introducing the intervention has led to an increase in Maggie's symbol production (especially considering that this behavior is not present prior to the intervention in two of the contexts). However, note that in the school and playground contexts the effect is delayed. The quantitative procedures can be applied to summarize the results and make them susceptible to meta-analysis. For the dollhouse and playground contexts, the GLS and the MPD yield very similar average difference, 1.34 and 1.40, respectively for dollhouse, and 3.22 and 2.95, respectively for

playground. This similarity is probably related to the fact that the correlation estimated by the Cochrane-Orcutt method in both cases is approximately .04 and thus the data transformation does not produce major differences. For the school context the MPD yields a value of 1.75 and the GLS 1.33, a difference that might be attributed to removing an estimated serial dependence of .33. If we summarize the results for all three baselines using the arithmetic mean (representing the average increase in symbols produced by Maggie in all three contexts after applying the aided language modeling), the results are very close: 1.96 according to the GLS and 2.03 as computed by the MPD.

The second data set (Figure 2) corresponds to a study carried out by Dolezal, Weber, Evavold, Wylie, and McLaughlin (2007) on middle school students aged 12-13 years and enrolled in a seventh grade reading classroom. The data were gathered for a participant called Jen, presenting difficulty in attending to task in reading, during two baseline phases (Direct Instruction *Corrective Reading program*) and two intervention phases in which a reinforcement package consisting of self-contracting, a token economy, and student graphing was added. The measurements in this ABAB design correspond to reading rates during two minute periods.

INSERT FIGURE 2 ABOUT HERE

Visual inspection suggests that after the first introduction of the reinforcement there is an increase in the behavior, which appears to be progressing. After the intervention is withdrawn, there is an immediate level change (a drop), but the behavior is seemingly improving afterwards even in absence of reinforcement (i.e., in the second A phase). Finally, when the intervention is re-inserted we observe an upward shift in level and the

behavior also seems to be more stable. Considering the whole data pattern, the degree of overlap between the measurements belonging to the two different conditions is rather small, but experimental control is not clearly demonstrated. For these data, we can use the quantitative procedures for helping us decide to what extent the improvements in the behavior may be related to the researchers' intervention.

When the design structure is ABAB it seems most meaningful to compare only adjacent phases (Gast & Spriggs, 2009). For the comparison between the first two phases, the GLS yields an increase of 37 words, and the MPD of 52 words. Nonetheless, the change after the reinforcement is withdrawn is not so clear, given that a negative level change is taking place alongside a positive slope change (i.e., on average there are less words read in the third phase but the increase in the behavior is more pronounced in this phase). Therefore, it might be logical to expect a "smaller" (however defined) drop in the behavior of interest: the MPD provides a value of −54, whereas the GLS yields a value of −120, although this latter large drop is not visible in the data. The last comparison between adjacent (third and fourth) phases was also problematic in terms of visual analysis, given that an increase in the target behavior was expected but not clearly identified. Both procedures concur that there is actually no increase when the third phase trend is taken into account: the GLS provides a value of −25.52 and the MPD −8.33. It is not straightforward to state which of the two values represents better the graphed data, although the impression is that there might actually be zero change in the trend after re-introducing the intervention.

For obtaining a single indicator for the whole design we used here the simple average of the two-phase comparisons, although weighting is also possible (Manolov & Solanas, 2012). It seems appropriate to give the same sign to the differences in the desired direction (i.e., an increase after introducing the intervention, a decrease after

withdrawing it, and another increase with the re-introduction). Using this logic, the averages obtained for the two procedures are 43.83 for the GLS and 32.56 for the MPD, representing the gains in terms of words read per 2 minutes associated with reinforcement (or the losses due to its withdrawal). Nevertheless, it should be kept in mind that the demonstration of the relationship between intervention and behavior does not meet current standards, as proposed by Kratochwill and colleagues (2010).

In the following section, we present the conditions in which the two procedures commented and illustrated above are compared. The simulation parameters used for specifying these conditions are made explicit in order to enhance transparency through full reporting and to make possible replications of the simulation.

Method

Data Generation

Different series lengths of two-phase (AB, baseline followed by intervention) designs were included in the current study. It was compulsory to include a condition with $n = 20$ data points for the whole series, given that it is the minimum recommended for the autocorrelation estimation in the GLS (Maggin et al., 2011). In the current study, it was specified as $n_A = n_B = 10$, for the baseline and treatment phases respectively. Shorter series were also included in order to represent better real data sets, specifically, $n_A = n_B = 5$ for ten-measurement series and $n_A = 5$, $n_B = 10$ for fifteen-measurement series. In the single-case studies review by Shadish and Sullivan (2011) it was found the average series lengths range, approximately, from 9 to 40, according to the psychological field, with a mean $n \approx 15$ for the single-case studies published in the *Journal of School Psychology*. In terms of the breadth of the experimental conditions simulated, the study

focuses only on two-phase AB designs, although it is also relevant for multiple baseline designs and any comparisons between two adjacent phases pertaining to different conditions (e.g., ABAB). The amount of phases per cases was found to be generally comprised between 2 and 4, with an average of 2.5 for single-case studies published in the *Journal of School Psychology* (Shadish & Sullivan, 2011).

Data were generated using a model proposed by Huitema and McKean (2000) expressed as $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot D_t + \beta_3 \cdot D_t \cdot [T_t - (n_A + 1)] + \varepsilon_t$, given its flexibility for representing a great variety of data patterns. This model allows specifying different types of general trend in data via the variable *T*, defined as the session number, taking all integer values between *1* and *n* for series with length *n* in order to represent linear trend. For the conditions with quadratic trend, *T* was set as all squared integer values from 1^2 to n^2. The presence of trend was marked here with $\beta_1 = .1$, whereas this parameter was set 0 for the conditions without trend. A level change was simulated through the dummy variable for change in phase: D_t was set to 0 for the first phase (i.e., there were n_A zeros) and D_t was set to 1 for the second phase (i.e., there were n_B ones). For simulating change in slope we used the interaction between the time variable and change in phase dummy variable ($D_t \cdot [T_t - (n_A+1)] = 0$ for the first phase and $D_t \cdot [T_t - (n_A + 1)]$ taking all integer values between 0 and $n_B - 1$ for the second one). The magnitudes of level change (β_2) studied were 1 and 3 and for slope change (β_3) were set to .5 and 1, given that this latter type of effective is progressive and produced a greater average difference between the measurements of the two phases.

There are several possibilities for defining the error term (ε_t), which allows expanding the experimental conditions. The main expression used was the one corresponding to the first-order autoregressive process, AR(1), $\varepsilon_t = \rho_1 \cdot \varepsilon_{t-1} + u_t$, assuming that each error term is a function of the immediately preceding error term plus

a random disturbance. This process is the one commonly used in simulation studies (e.g., Ferron & Onghena, 1996; Levin et al., 2011), and has been considered the most reasonable for psychological data (Glass et al., 1975). For this process ρ_1 was set to −.3, 0 (white noise error), .3, and .6. These values cover the range of typical negative and positive autocorrelation as reported by Parker (2006) and most of the range reported by Shadish and Sullivan (2011), especially considering the negative bias in the autocorrelation estimator the latter authors used. The expression was also used for specifying four combinations of phase-specific degrees of serial dependence (different degrees of serial dependence have already been reported for psychological data; Robey et al., 1999): ρ_{1A} = −.3 and ρ_{1B} = .3, ρ_{1A} = 0 and ρ_{1B} = .3, ρ_{1A} = 0 and ρ_{1B} = .6, and ρ_{1A} = .3 and ρ_{1B} = −.3. In some conditions, only the baseline phase was generated via an AR(1) process, whereas for the intervention phase the first-order moving average, MA(1), model $\varepsilon_t = u_t - \theta_1 \cdot u_{t-1}$ was used, or vice versa. The MA(1) has also been suggested for behavioral data (Harrop & Velicer, 1985) and it assumes that each error term is a function of the current and the pervious random disturbances, u_t and u_{t-1}. Thus, in these conditions not only the degree of autocorrelation, but also the autocorrelation scheme was different between the two phases. The specific combinations tested were: ρ_{1A} = .3 and θ_{1B} = .33, ρ_{1A} = −.3 and θ_{1B} = −.33, θ_{1A} = −.33 and ρ_{1B} = −.3, and θ_{1A} = .33 and ρ_{1B} = .3. Keep in mind that a negative sign of the θ_1 parameter implies positive serial dependence and that θ_1 = −.33 corresponds to $\rho_1 \approx .298$ (McCleary & Hay, 1980).

Regarding the disturbances (u_t), these were specified to follow three distinct distributions: the normal, a more platykurtic one (the uniform), and a more skewed one (the negative exponential). For all distributions the mean was set to zero and for the homoscedastic conditions the standard deviation (*SD*) to one, in order to make the data patterns comparable. In the conditions of unequal variances, the cases in which

variability in phase B is twice and four times the variability in phase A, in terms of standard deviation, were studied (i.e., SD_B = 2 and SD_B = 4 with SD_A = 1 in all conditions).

Finally, for each of the data patterns specified using the abovementioned expressions 10,000 samples were generated using R (R Development Core Team, 2012). Thus, for each condition a distribution of 10,000 values was obtained for the GLS and the MPD and in order to compare the procedures the mean of each distribution was used. Regression analyses were carried out using functions already implemented in R, whereas the *lmtest* (Zeileis & Hothorn, 2002) package was used for performing the Durbin-Watson test.

Data Analysis

In the present study evidence was obtained according to three different criteria: distortion due to confounding factors (serial dependence, trend, and inequality of variance) in absence of intervention effect, detection of the presence of intervention effect, and efficiency of the indicators in all conditions tested.

The distortion due to confounding factors is made in the conditions without intervention effect. When there is no intervention effect programmed, the desired value of the difference between phases is zero, which would be indicative of precisely estimated (null) effect and insensitivity to the confounding factors. Special attention is paid to values which are outside the interval ranging from −.10 to .10. This arbitrary reference was chosen given that it represents a first decimal point deviation from zero.

The detection of existing effects is studied in conditions in which intervention effect was simulated. The detection of effects was studied in data patterns with no systematic elements programmed (e.g., trend, serial dependence), apart from the level or slope change itself. In this case, as a general rule, the greater values yielded by the procedures are more desirable, although we will refine this criterion in the corresponding subsection of the Results.

Finally, in order to gather information on the efficiency of the indicators as estimators of the effect programmed (zero or greater than zero), the sampling distribution was estimated for each condition studied. The standard error was estimated as the standard deviation of these sampling distributions, that is, as a quantification of the variability of the estimator (GLS or MPD) about the point estimate of the parameter. Note that greater indicator variability (i.e., greater standard error) means less efficient performance of the estimator.

Results

Absence of Intervention Effect: Distortion Due to Confounding Factors

When a constant non-null degree of serial dependence ($\rho_1 \neq 0$) is present in the data, both procedures yield values close to the desired zero. These values are $\leq |.05|$ in all cases, as Table 1 shows. When the degree or the process of serial dependence are phase-specific (Tables 2 and 3), the effect size quantifications provided by the procedures do not deviate excessively from zero either.

INSERT TABLES 1, 2, AND 3 ABOUT HERE

Columns three to six in Table 4 contain the results for the distortion due to trend in the GLS and the MPD values. Considering the fact that the MPD estimates and projects linear baseline trend, it was expected to find that it is not distorted by linear trend, whereas quadratic trend does prove to be problematic for the procedure. For the GLS, it can be seen that when only linear trend is taken into account, quadratic trend is problematic. However, when step 1 of the GLS (as presented in "The procedures described" section) included a time variable for modeling quadratic trend, the GLS remained unaffected and thus performed adequately, although the estimator lost efficiency (the results are not presented here).

Regarding the effect of heteroscedasticity, as the last four columns of Table 4 suggest, the general finding is that both procedures are unaffected by the presence of unequal data variability in the two phases. Thus, even when the standard deviation in the intervention phase is four times as large as in the baseline phase, this does not lead to a systematic overestimation or underestimation for any of the two procedures.

INSERT TABLE 4 ABOUT HERE

Presence of Intervention Effects: Detection of the Effects

Given that the detection of effects was studied in the context of homoscedastic data with *SD* = 1, a level change equal to 1 is equivalent to a standardized mean difference (δ) of

1. Consequently, when the β_2 parameter was set to 3, δ is equal to 3. Hence, the results show that in these cases the MPD estimates these quantities with no systematic bias, whereas the values provided by the GLS are lower, as shown on Table 5. The difference between the procedures is greater for series with $n \geq 15$. When the detection of slope change was the focus of attention, it is necessary to compare the values yielded by the indices to mean difference between phases expected according to the value specified for β_3 and also according to the length of the treatment phase. Given that β_3 is actually the difference between two successive points in the treatment phase and it is getting accumulated with each consecutive measurement, the average change between phases can be expressed as $\sum_{i=0}^{n_B-1} i\beta_3 \Big/ n_B$. Thus, when β_3 = .5, the average difference between phases is equal to 1 for n_B = 5 and it is equal to 2.25 for n_B = 10. When β_3 = 1, the mean difference between the phases is twice these values[2]. Once again, the quantifications provided by the MPD are closer to the simulated ones and greater than the ones yielded by the GLS. Therefore, the MPD seems to be more sensitive to intervention effects than the GLS. The sensitivity difference for slope change is greatest for n = 20, with the MPD providing twice as large values as the GLS.

INSERT TABLE 5 ABOUT HERE

Efficiency of the Procedures: Standard Error

[2] Note that these average differences are actually standardized given that SD=1.

The standard errors for both procedures are presented in the five tables, according to the condition simulated. The main result that ought to be commented is that the standard errors are greater for the MPD than for the GLS in all cases. In order to be able to assess the difference between the techniques, we offer the ratios of the standard error of the MPD as compared to the GLS. For independent data series these ratios range from 1.17 to 1.50 with an average (for all series lengths and random disturbance distributions) of 1.28 and for high positive autocorrelation from 1.18 to 1.64 with an average of 1.36. The techniques are closer in performance (an average ratio of 1.13 for AR(1) and 1.15 for MA(1)) when there is positive autocorrelation in the baseline and negative autocorrelation in the intervention phase. Complementarily, the GLS is much more efficient (an average ratio of 1.52 for AR(1) and 1.57 for MA(1)) in the reversed condition. Given that heteroscedasticity is associated with higher standard errors for both procedures, the relative difference between them is reduced (an average ratio of 1.11 when $SD_{treatment} = 4 \cdot SD_{baseline}$). The loss of efficiency present when using the simpler procedure is especially evident when a slope change is programmed (and average ratio of 1.57 for $\beta_3 = 1$). Finally, two results common for both techniques should be commented. First, the standard error is greater for positive autocorrelation. Second, the variability in the indicators' values was related to series and phase lengths. The conditions in which both procedures showed less efficient performance were when $n_A = 5$, $n_B = 10$ followed by $n_A = n_B = 5$. Thus, not only the shortness of the baseline is detrimental, but also (actually to a greater degree) the lack of balance between the phase lengths.

Discussion

The present study focused on two quantitative procedures for single-case data analysis in the search for appropriate indicators that can be used to enhance establishing the evidence base for psychological interventions (Wampold, Goodheart, & Levant, 2007). In the present section a summary of the results is presented before discussing in the Conclusions the practical implications of the results of the current study taken together with previous findings on how single-case data can be analyzed numerically.

The main differences between the procedures are in terms of standard error and sensitivity, given that the distortion due to confounding factors (serial dependence, linear trend, heteroscedasticity) is minimal in both cases. In fact, the GLS proved to function well even beyond the conditions that Maggin et al. (2011) recommend for the autocorrelation estimation step (i.e., $n \geq 20$ and when the same degree and the serial dependence scheme is present for the two phases, plus regression's assumption of normal homoscedastic error term). Thus, this regression-based procedure seems to outperform the previously tested ones – the three types of analysis presented by Gorsuch (1983), the Allison and Gorman (1993) model, and White et al.'s (1989) index, although here the unstandardized version of the GLS indicator was tested.

In terms of standard error, the GLS is a more efficient estimator of the change between phases, which is logical given the characteristics of the procedure. Technically, a more efficient estimator is one that shows lower variability around the parameter when estimates are obtained from samples drawn from the same population. The practical implications are that if the estimator is more efficient, as is the GLS, then the researcher can have greater confidence that the numerical value yielded for his/her data is close to the true value. In this study, the main question was how well a simpler procedure (the MPD) would match the performance of the GLS in terms of efficiency. In that sense, the GLS is especially preferable for estimating slope change and for cases

in which there is evidence of negative serial dependence in the baseline and positive serial dependence in the treatment phase. For the remaining conditions the difference does not seem to be as large. It has to be mentioned that both procedures showed worse performance in terms of efficiency when the treatment phase was longer than the baseline phase (i.e., when n_A = 5 and n_B = 10 here). Given that Shadish and Sullivan (2011) found that 45% of the baselines had fewer than 5 data points, it is logical to expect longer treatment phases, which appears to be detrimental for the functioning of the procedures. Therefore, longer baselines are necessary not only for ensuring stable behavior and a solid basis for comparison, but also for favoring the quality of the studies' quantification.

In terms of detection of existing intervention effects, the MPD was found to be more sensitive yielding greater values than the GLS when both were compared in the same conditions. That is, using the parallelism with statistical significance testing, the MPD can be said to be more powerful (always keeping in mind that we actually did not estimate the proportion of times a false null hypothesis is rejected). Regarding detection of effects, one result should be pointed out – the fact that when simulating the same parameter β_3, both procedures yielded different values according to series length. A similar limitation has been shown for the Allison and Gorman (1993) model and for White et al.'s (1989) index by Van den Noortgate and Onghena (2003). This finding was expected as the indices quantify mean difference between phases and that a treatment phase slope implies a nonstationary process (i.e., series in which the average level changes systematically with each successive measurement obtained). Thus, the average difference becomes larger (i.e., the experimental unit continues improving) with time and so the indicators ought to reflect such an increasing difference with respect to the initial situation, that is, the baseline level. This feature of the performance of the

techniques makes evident that a slope change should not be quantified via an indicator focusing on the average levels in the phases compared, given that the result could not be unambiguously attributed to the magnitude of effect, as series length is also relevant. This may be especially problematic when summary measures are to be meta-analyzed, as it may be necessary to include series length as control variable.

Finally, it has to be considered how the proposed MPD can be included in meta-analyses of single-case studies. Weighting studies is an important part of the quantitative integrations (Hedges & Olkin, 1985) and an especially recommended alternative is to use the standard error of the primary indicator (Whitlock, 2005). However, the standard error is not known for all indicators. Currently there is a proposal for making such kind of information available after estimating the relevant sampling distribution (Manolov & Solanas, 2012), but more discussion on this issue is necessary.

Conclusions

For establishing the evidence base of psychological practice it is important that applied researchers complement their substantive (e.g., educational) criteria and the visual analysis of the graphed single-case data with some numerical indicators of the degree of behavioral change. Such quantitative results can be used for accountability, communication between researchers, and also for future meta-analyses. Therefore, it is important to identify suitable analytical procedures. Among the criteria that can be used for assessing whether a procedure is suitable or not we can point tentatively at the interpretability of the information provided by the numerical indicators and the ease of obtaining these numerical values by hand calculation or statistical software. Regarding

these two aspects, the Institute of Education Sciences (2012) favors the development of practical techniques that can be used by most researchers and do not require advanced statistical knowledge. Another relevant aspect related to the use of the indicator criteria are the design requirements and data assumptions necessary. Finally, the evidence on the performance of the techniques should be considered. We refer to this last aspect, which was the aim of the current study, as statistical soundness.

Table 6 presents a summary of the strengths and limitations of the two procedures tested here, as well as for other techniques mentioned in the introductory section, integrating the current study results with previous findings (e.g., Ferron & Onghena, 1996; Manolov et al., 2011; Parker & Brossart, 2003). We hope that this summary table might help researchers choose a technique for analyzing single-case data. As a tentative recommendation to researchers, we emphasize the usefulness of the NAP for quantifying data overlap and converting it to Pearson's Phi, when the data do not present trend and the measurements of different phases share at least some values. Among the easy to obtain and interpret procedures, the SLC is more useful when the purpose is to estimate separately level and slope change or when data require this (e.g., the effects are in opposite directions), whereas the MPD can be used to obtain a single quantification, but keeping in mind that the quantification of a slope change is not independent from series length. The more complex GLS can be used in lieu of the MPD for obtaining more efficient estimates of average phase differences. Another advantage of this procedure is the possibility to take into account quadratic trend. GLS offers also the possibility to standardize the effect size, whereas for the MPD we have proposed a tentative solution for taking into account the variability in the data.

INSERT TABLE 6 ABOUT HERE

Finally, it should be noted that we do not state that quantification is either the only or the most appropriate way to assess intervention effectiveness, given the importance professionals' knowledge of the client and the specific problem treated and the need for visual inspection of graphed data to ascertain the functional relationship between the intervention and the target behavior. Future research may focus on the preferences of applied researchers regarding the quantitative procedures they actually use and the ones they would be willing to apply, considering Parker et al.'s (2006) claim that "to be acceptable to most single-case intervention practitioners, research tools must work with and complement visual analysis of data plots" (p. 419).

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31,* 621-631.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd Ed.). Boston, MA: Pearson.

Brossart, D. F., Parker, R. I., & Castillo, L. G. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods, 43,* 710-719.

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,* 531-563.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41,* 477-485.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Colón, M. (2006). Improving the reliability and validity of visual inspection of data by behavior analysts: An empirical comparison of two training methods to improve visual inspection and interpretation, the job aid and the conservative dual-criteria. Unpublished doctoral dissertation, Florida State University.

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

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

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

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intersubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Dolezal, D. N., Weber, K. P., Evavold, J. J., Wylie, J., & McLaughlin, T. F. (2007). The effects of a reinforcement package for on-task and reading behavior with at-risk and middle school students with disabilities. *Child & Family Behavior Therapy, 29*, 9-25.

Drager, K. D. R., Postal, V. J., Carrolus, L., Castellano, M., Gagliano, C., & Glynn, J. (2006). The effect of aided language modeling on symbol comprehension and production in 2 preschoolers with autism. *American Journal of Speech – Language Pathology, 15*, 112-125.

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

Ferron, J. M., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64*, 231-239.

Ferron, J. M., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education, 70*, 165-178.

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.

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

Gast, D. L., & Spriggs, A. D. (2009). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199−233). London: Routledge.

Glass, G. V., Willson, V. K., & Gottman, J. M. (1975). *The design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Greenwald, A. G. (1976). Within-subject designs: To use or not to use? *Psychological Bulletin, 8*, 314-320.

Grissom, R. J., & Kim, J. J. (2012) *Effect size for research: Univariate and multivariate applications* (2nd ed.). London: Routledge.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30*, 313-326.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

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.

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

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38-58.

Institute of Education Sciences. (2012). Request for applications: Statistical and research methodology in education. Retrieved July 12, 2012 from http://ies.ed.gov/funding/pdf/2013_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. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Lawrence Erlbaum.

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.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Knapp, T. J. (1983). Behavioral analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155-164.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2*, 291-307.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: 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.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th Ed.). London: McGraw-Hill.

Lall, V. F., & Levin, J. R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests. *Journal of School Psychology, 42*, 61-86.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012, June 6). Nonparametric statistical tests for single-case systematic and randomized ABAB…AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*. Advance online publication. doi: 10.10167j.jsp.2012.05.001

Levin, J. R., Lall, V. F., & Kratochwill, T. R. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology, 49*, 55-79.

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*, 301-321.

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860-875.

Manolov, R., & Solanas, A. (2012, July 16). Assigning and combining probabilities in single-case studies. *Psychological Methods*. Advance online publication. doi: 10.1037/a0029248

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*, 533-545.

McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. Beverly Hills, CA: Sage.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 3*, 87-101.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.

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

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*, 783-786.

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

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326-338.

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

Parker, R. I., & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

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

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

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single-subject clinical outcome research: Designs, data, effect sizes, and analysis. *Aphasiology, 13*, 445-473.

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*, 163-187.

Schneider, N., Goldstein, H., & Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment and Intervention, 2*, 152-162.

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

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.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140-160.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84*, 489-502.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners judgements of intervention effectiveness. *Journal of Special Education, 22*, 433-446.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

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

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

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation, 18*, 385-401.

Tate, R. L., Togher, L., Perdices, M., McDonald, S., & Rosenkoetter, U. (2012, July). *Developing reporting guidelines for single-case experimental designs: the SCRIBE project*. Paper presented at the 9th Conference of the Neuropsychological rehabilitation special interest group of the World federation for neurorehabilitation, Bergen, Norway.

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*, 423-429.

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.

Wampold, B. E., Goodheart, C., & Levant, R. (2007). Clarification and elaboration on evidence-based practice in psychology. *American Psychologist, 62*, 616-618.

White, O. R. (1972). *The split-middle: A quickie method of trend analysis*. Eugene, OR: Regional Center for Handicapped Children.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology, 18*, 1368−1373.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7-10. URL http://CRAN.R-project.org/doc/Rnews/

Appendix

R code for computing the MPD. The lines preceded with # indicate comments which explain the sentences listed below. The user has to put the measurements separated by commas in the first two lines replacing the example, for the baseline and the treatment data, respectively. The code is copy-pasted into the R console and offers the MPD value and a graphical representation of the actual and the predicted data.

```
# User input
baseline <- c(1,2,3)
treatment <- c(5,6,7)

# Obtain phase length
n_a <- length(baseline)
n_b <- length(treatment)

# Estimate baseline trend
base_diff <- rep(0,(n_a-1))
for (i in 1:(n_a-1))
   base_diff[i] <- baseline[i+1] - baseline[i]
trendA <- mean(base_diff)

# Project baseline trend to the treatment phase
treatment_pred <- rep(0,n_b)
for (i in 1:n_b)
   treatment_pred[i] <- baseline[1] + trendA*(n_a+i-1)

# Compare the predicted and the actually obtained treatment data and obtain the average difference
mpd <- mean(treatment-treatment_pred)
print(mpd)

# Represent graphically the actual data and the projected treatment data
info <- c(baseline,treatment)
info_pred <- c(baseline,treatment_pred)
minimo <- min(info,info_pred)
maximo <- max(info,info_pred)
time <- c(1:length(info))
plot(time,info, xlim=c(1,length(info)), ylim=c((minimo-1),(maximo+1)), xlab="Measurement time", ylab="Behavior
of interest", font.lab=2)
abline(v=(n_a+0.5))
lines(time[1:n_a],info[1:n_a])
lines(time[(n_a+1):length(info)],info[(n_a+1):length(info)])
axis(side=1, at=seq(0,length(info),1),labels=TRUE, font=2)
points(time, info, pch=24, bg="black")
points (time, info_pred, pch=19)
title (main="Predicted (circle) vs. actual (triangle) data")
```