# Comparing N = 1 effect size indices in presence of autocorrelation

Rumen Manolov and Antonio Solanas

Department of Behavioral Sciences Methods,

Faculty of Psychology,

University of Barcelona

## ABSTRACT

Generalization from single-case designs can be achieved by means of replicating individual studies across different experimental units and settings. When replications are available, their findings can be summarized using effect size measurements and integrated through meta-analyses. Several procedures are available for quantifying the magnitude of treatment's effect in N = 1 designs and some of them are studied in the current paper.

Monte Carlo simulations were employed to generate different data patterns (trend, level change, slope change). The experimental conditions simulated were defined by the degrees of serial dependence and phases' length. Out of all the effect size indices studied, the Percent of nonoverlapping data and standardized mean difference proved to be less affected by autocorrelation and perform better for shorter data series. The regression-based procedures proposed specifically for single-case designs did not differentiate between data patterns as well as simpler indices.

**Key words:** single-case, AB designs, effect size, autocorrelation

N = 1 designs have been criticized due to the problematic statistical generalizations. A possible solution of this problem consists in replicating across subjects and settings in order to establish the generality of the treatment effects. The quantitative integration of these replications can be accomplished by means of meta-analysis. A prior step to integration is summarizing the evidence from each study, a stage in which effect sizes are of maximum relevance. The measurements of the magnitude of effect have gained importance as they overcome p-values' limitations (Cohen 1990; 1994; Kirk, 1996; Rosnow & Rosenthal, 1989; Wilkinson & The Task Force on Statistical Inference, 1999). Effect size is an objective measurement of the strength of the intervention and provides clinical and social researchers more useful information than the significance level. In contrast with the latter, effect size are not systematically affected by sample size (Parker & Brossart, 2003) and focuses on the strength of association between the independent and the dependent variables, instead of centering on the null hypothesis (Kromrey & Foster-Johnson, 1996). Moreover, effect size allows comparing treatments and is useful for documenting results for posterior meta-analysis and power analysis (Parker & Hagan-Burke, 2007). Another advantage is the possibility to construct confidence intervals about the effect size (Kirk, 1996).

One of the peculiarities of single-case designs is that they generally include few measurement times (Huitema, 1985). On the other hand, several surveys (e.g., Busk & Marascuilo, 1988; Matyas & Greenwood, 1991; 1996; Parker, 2006) report that autocorrelation is a common feature of N = 1 designs. It has been claimed that even low and statistically non-significant levels of autocorrelation can have critical influence on the analytical techniques employed (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Moreover, empirical findings suggest that autocorrelation affects a great variety of statistical techniques like ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983), the binomial test and the split-middle method (Crosbie, 1987), randomization tests (Gorman & Allison, 1996; Sierra, Solanas, & Quera, 2005) and also visual analysis (Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990).

The typical phase length and the likely presence of serial dependence have influenced the lack of consensus about the optimal effect size measurement in single-case research. The most frequent formulae such as standardized mean differences (e.g., Cohen's *d*; Hedges' *g*; Glass' Δ) and correlations (e.g., $\eta^2$; $\omega^2$; $R^2$), have been conceptualized and developed for group designs and focus solely on the average level in the control and treatment conditions. There have also been proposed indices destined specifically to N-of-1 designs, such as the Percent of Nonoverlapping Data (PND) or the regression indices (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-86; Gorsuch, 1983; White, Rusch, Kazdin, & Hartmann, 1989). PND, as its name suggests, centers on a criterion frequently used in visual inspection, which is still the most commonly applied single-case data analysis technique (Parker, Cryer, & Byrns, 2006). The regression procedures take into account mean levels and the possible slope changes between conditions and also control for trends not associated with the intervention. The comparison between studies is enhanced by the possibility of converting one type of index into another (Friedman, 1982).

Each of the indices mentioned has its drawbacks: deficient performance in presence of outliers and trend, ignoring all phase A data points but one (PND); no account for changes in slope (Gorsuch's Trend analysis and White et al.'s *d*); conservativeness, attainment of more than one magnitude of effect index and impossibility to obtain a negative *d* (Center, Skiba, & Casey's procedure); possibility to produce unreliable estimates of trend due to short baseline and overestimation of effect size (Allison & Gorman's procedure). Regarding the limitations of the latter, which appears to be the conceptually most appropriate one, too large effect sizes may potentially affect interpretability (Campbell, 2004). With respect to that, Scruggs & Mastropieri (1998) point out that an effect size of $d$ = 3.0 implies that percentile 50 of the treatment phase corresponds to percentile 99.9 of the baseline phase, making greater values of *d* practically useless. Finally, applied researchers have to keep in mind that when the parametric assumptions of regression-based procedures are not met the correctness of the effect sizes calculated is not guaranteed. We performed a small revision of scientific literature and found that PND seems to be employed more frequently (e.g., Bellini, Peters, Benner, & Hopf, 2007; Mathur, Kavale, Quinn, Forness, & Rutherfod, 1998; Scruggs & Mastropieri, 1994; Scruggs, Mastropieri, Forness, & Kavale, 1988) than regression-based methods (Allison, Faith, & Franklin, 1995; Skiba, Casey, & Center, 1986), probably due to the relatively greater complexity of the latter.

The objective of the present investigation was to assess the performance of six proposed measures of effect sizes for AB designs in presence of different degrees of autocorrelation. The comparison between the indices was done in terms of $R^2$ (except for PND) due to the fact that this indicator ranges from 0 to 1 and is easily interpreted as “the variance of the dependent variable explained by the change in phase”. Due to the fact that estimating autocorrelation from real data, and testing it for significance, may be problematic (Huitema & McKean, 1991; Matyas & Greenwood, 1991), we decided to test the effect size procedures with data constructed with known parameters (i.e., serial dependence, trend, level change, slope change), a method that has already been applied in single-case effect size studies (Parker & Brossart, 2003). Another aim was to evaluate the influence of series length, as suggested by Campbell (2004).

# Method

*Design selection*

Two-phase AB designs with different total (N) and phase length ($n_A$ and $n_B$) were studied.  Short series were chosen as they are more feasible in applied settings: a) N = 10; $n_A$ = $n_B$ = 5. b) N = 15; $n_A$ = 5; $n_B$ = 10. c) N = 15;

$n_A = 7$; $n_B = 8$. d) N = 20; $n_A = 5$; $n_B = 15$. e) N = 20, $n_A = n_B = 10$. f) N = 30, $n_A = n_B = 15$.

*Data generation*

The data for the abovementioned series lengths were generated according to an expression that allows specifying level and slope changes, and trend. The statistical model was the same as in previous investigations (e.g., Huitema & McKean, 2000; 2007):

$y_t = \beta_0 + \beta_1 * T_t + \beta_2 * D_t + \beta_3 * SC_t + \varepsilon_t$, where:

$y_t$: the value of the dependent variable at moment *t*;

$\beta_0$: intercept;

$\beta_1, \beta_2, \beta_3$: partial correlation coefficients;

$T_t$: value of the time variable at moment *t* (takes values from 1 to N);

$D_t$: dummy variable for level change (0 for phase A and 1 for phase B);

$SC_t$: value of the slope change variable. $SC_t = [T_t – (n_A + 1)]*D_t$. Takes 0 for phase A, and values from 0 to $(n_B − 1)$ for phase B.

$\varepsilon_t$: error term;

The error term ($\varepsilon_t$) was generated following a first-order autoregressive model: $\varepsilon_t = \varphi_1 * \varepsilon_{t–1} + u_t$. The values of serial dependence ($\varphi_1$) ranged from –0.9 to 0.9 in steps of 0.1. The $u_t$ term represents white noise at moment *t* and $\varepsilon_1 = u_1$.

The value of the intercept parameter $\beta_0$ was set to zero as it does not affect effect size calculation. On the other hand, our goal was to guarantee suitable comparisons between experimental conditions. Therefore, it was important that the two types of effects (i.e., level change associated with parameter $\beta_2$, and slope change associated with $\beta_3$) and trend (extraneous variable associated with parameter $\beta_1$) produced comparable mean differences between phase B and phase A. Firstly, two criteria were chosen: a) *series length*: the shortest design was chosen $n_A = n_B = 5$ in order to explore if longer series imply better effects detection; b) *the partial correlation coefficient*: level change ($\beta_2$) was selected as it maintains constant throughout the whole intervention phase. As the $u_t$ term was generated following $\mathbf{N}(0,1)$, the phase A values approximate zero ($y_{Ai} \approx 0$). Being present a level change of $\beta_2$, $y_{Bi} = y_{Ai} + \beta_2 = 0 + \beta_2 = \beta_2$. $\beta_2 = 0.3$ was chosen as it proved to avoid floor and ceiling effects (i.e., $R^2$ not approaching 0 nor 1, respectively). The change in slope produces ($n_B - 1$) increments and it was necessary to find a $\beta_3$ value so that the median phase B point be equal to $\beta_2$, which will make the phase B mean also equal to $\beta_2$. As $\bar{y}_B - \bar{y}_A = \beta_2$, a $\beta_3$ value implying the same mean difference can be calculated as

$\beta_3 = \dfrac{\beta_2}{\frac{n_B - 1}{2}}$, which for $\beta_2 = 0.3$ leads to $\beta_3 = \dfrac{0.3}{\frac{5-1}{2}} = \dfrac{0.6}{4} = 0.15$

As trend involves increments from the first observation, the accomplishment of the $\bar{y}_B - \bar{y}_A = \beta_2$ criterion required meeting the following equality $y_{Bi} - y_{Ai} = \beta_2$. The needed $\beta_1$ value can be found as

$\beta_1 = \frac{\beta_2}{\frac{n_A + n_B}{2}}$, which for $\beta_2 = 0.3$ leads to $\beta_1 = \frac{0.3}{\frac{5+5}{2}} = \frac{0.6}{10} = 0.06$

We could verify that the $\beta_1$ and $\beta_3$ values are appropriate for producing $\beta_2$ mean differences even for the most extreme levels of serial dependence (−0.9 and 0.9), whenever $n_A = n_B$. In total there were eight data patterns studied, defined by the presence and combination of trend, level change, and slope change (i.e., $\beta_1$, $\beta_2$, and $\beta_3$ being equal to or different from zero).

It is likely that for series with high negative autocorrelation unstable baselines be obtained. Therefore, we used a large number of iterations in order to ensure that the indices' comparison does not depend on few clinically improbable data sets.

The 50 number previous to each simulated data series were eliminated in order to reduce artificial effects (Greenwood & Matyas, 1990) and to avoid dependence between successive data series (Huitema, McKean, & McKnight, 1999).

### *Analysis*

We calculated the effect size for each experimental condition using the following indices:

#### Percent of Nonoverlapping Data

1) Calculate the number of phase B data points that exceed the highest data point in phase A. Simulating increases in behavior with the introduction of treatment ensures that this step is appropriate.
2) Divide the value obtained in step 1 by the number of observations in phase B and multiply by 100 in order to convert the proportion in percentage.

Cohen's *d*

1) Obtain the difference between the means of both phases: $\bar{y}_B - \bar{y}_A$.
2) Calculate the standard deviation of each phase.
3) Divide the value obtained in step 1 by the phase A standard deviation or by the pooled standard deviation (obtaining $d_A$ and $d_{AB}$, respectively).
4) Convert *d* to $R^2$, using $R^2 = \frac{d^2}{d^2+4}$ .

Gorsuch's (1983) Trend analysis:

1) Calculate a simple linear regression using time (T = 1, 2, …, n) as a predictor variable and the original dependent variable: $\mathbf{Y = a + b_t{*}T + u_t}$
2) Calculate a simple linear regression using the treatment variable (X = 0 for phase A and X = 1 for phase B) as a predictor and the residual of the step 1 regression as a dependent variable: **residual(Y) = a +** $\mathbf{b_x{*}X + u_t}$
3) Calculate $R^2$ as the sum of squares explained by the step 2 model divided by the total sum of squares.

White et al.'s *d* (1989, using the correction in Faith, Allison, & Gorman, 1996)

1) Calculate a simple linear regression using phase A data and the time variable as predictor.

2) Use the step 1 regression coefficients (intercept and slope) to obtain the predicted value of the dependent variable for the last day of the B phase – this value is called $y_A$.

3) Calculate a simple linear regression using phase B data and the time variable as predictor.

4) Use the step 3 regression coefficients (intercept and slope) to obtain the predicted value of the dependent variable for the last day of the B phase – this value is called $y_B$.

5) Calculate the difference $y_B - y_A$ which represents the numerator in White et al.'s (1989) formula.

6) Calculate the pooled standard deviation of phases A and B.

7) Calculate the Pearson product-moment correlation coefficient between the dependent variable and the time variable.

8) Calculate *d* through the expression $d = \frac{y_B - y_A}{\sqrt{(1-r^2)*\sqrt{(s_A^2+s_B^2)/2}}}$.

9) Convert *d* to $R^2$.

Allison & Gorman (1993).

1) Calculate a simple linear regression using phase A data and the time variable as predictor.: $\mathbf{Y_A = b_0 + b_1 * T_A + e}$

2) Calculate the predicted values for Y and the residuals for both phases.

3) Calculate zero-order correlations between the treatment variable X (X = 0 for phase A and X = 1 for phase B) and residual(Y), on one hand, and between X*T and residual(Y), on the other. If both correlations share the same sign, then proceed with step 4. Otherwise, go to step 6.
4) Calculate a multiple linear regression with the treatment variable X and the X*T as predictors: **residual(Y) = $\mathbf{b_0}$ + $\mathbf{b_1}$*X + $\mathbf{b_3}$*X*T + e**
5) Obtain the adjusted $R^2$ for the step 4 equation.
6) In case the zero-order correlations associated with level and slope have different signs, it is only necessary to estimate the effect of the treatment variable X through a simple linear regression, as the change in slope will attenuate this effect. Obtain the adjusted $R^2$.

*Simulation*

The specific steps that were implemented in the Fortran programs (one for each of the six series length) were the following ones:

1) Systematic selection of each of the 19 degrees of serial dependence.
2) Systematic selection of the ($\beta_1$, $\beta_2$, and $\beta_3$) parameters for data generation: $2^3$ = 8 data patterns – autoregressive model; trend; level change; slope change; trend and level change; trend and slope change; level and slope change; trend, level and slope change.
3) 100,000 iterations of steps 4 through 17.
4) Generate an array with 50+N data following a normal distribution with mean zero and unitary standard deviation by means of NAG*fl90*

mathematical-statistical libraries (specifically external subroutines *nag_rand_seed_set* and *nag_rand_normal*).

5) Eliminate the first 50 numbers.

6) Assign the following N numbers to array $u_t$.

7) Establish $\varepsilon_1 = u_1$.

8) Obtain the array of $\varepsilon_t$ using the equation $\varepsilon_t = \varphi_1 * \varepsilon_{t-1}$.

9) Obtain the time array $T_t = 1, 2, \ldots, \mathrm{N}$.

10) Obtain the dummy treatment variable array $D_t$, where $D_t = 0$ for phase A and $D_t = 1$ for phase B.

11) Obtain the slope change array according to Huitema & McKean's (2007) expression: $SC_t = [T_t - (n_A + 1)]*D_t$ used for data generation.

11) Obtain the slope change array $T_t*D_t$ according to Allison & Gorman's (1993) procedure used in the effect size computation.

12) Obtain the $y_t$ array containing measurements (i.e., dependent variable) following Huitema & McKean's (2007) model: $y_t = \beta_0 + \beta_1*T_t + \beta_2*D_t + \beta_3*SC_t + \varepsilon_t$.

13) Calculate the Percent of Nonoverlapping Data.

14) Calculate effect size according to the two versions of Cohen's $d$ ($d_A$ and $d_{AB}$). Convert $d$ to $R^2$.

15) Calculate effect size ($R^2$) according to Gorsuch's (1983) Trend analysis.

16) Calculate White et al.'s (1989) $d$ and convert to $R^2$.

17) Calculate effect size (adjusted $R^2$) according to Allison & Gorman's (1993) procedure. NAG*fl90* libraries external subroutine

*nag_mult_lin_reg* was used to obtain the multiple regression coefficients.

18) Average the obtained $R^2$ from the 100,000 replications of each experimental condition.

During program elaboration the appropriate performance of the programs was verified through comparisons with the output of statistical packages and with the examples presented in Faith, Allison, & Gorman (1996).

## Results

Due to the low magnitude of effect estimates produced by Gorsuch's (1983) Trend analysis, this procedure will not be commented in the following sections. The values, ranging from 0.01 to 0.06 for all experimental conditions and concurring with Parker & Brossart's (2003) results, show the influence of autocorrelation and the zero sensitivity to the differential data patterns.

### *Autocorrelation effect*

To explore the effect produced by the presence of serial dependence in data, we constructed figures crossing each of the six effect size indices with the eight data patterns. In each of these 6 * 8 = 48 figures the degree of autocorrelation is placed on the abscissa and the index value ($R^2$ or percentage) on the ordinate, superimposing the different phase lengths. Visual inspection for simpler data patterns (i.e., when none or only one type of effect is present) showed that negative serial dependence is associated with lower $R^2$ values, while positive one correlates with higher effect size estimates. There appears to be an approximately linear relation between $\varphi_1$ and $R^2$. Figure 1 compares several techniques and illustrates the fact that for Cohen's *d* we observed a greater increment in $R^2$ for positive ($0.0 \leq \varphi_1 \leq 0.9$) than for negative autocorrelation ($-0.9 \leq \varphi_1 \leq 0.0$). As Figure 2 shows, for PND there is a nonlinear relation between autocorrelation and the effect size measurement which in this case, due to the peculiarities of the index, is the percentage itself rather than an $R^2$.

INSERT FIGURE 1 ABOUT HERE

INSERT FIGURE 2 ABOUT HERE

Comparing the differences in $R^2$ between high negative ($\varphi_1 = -0.9$) and zero autocorrelation, on one hand, and high positive ($\varphi_1 = 0.9$) and zero autocorrelation, on the other, it appears that White et al.'s *d* and Allison & Gorman's procedure are the most affected ones, while Cohen's *d* and PND are less sensitive to serial dependence. When the data pattern is more complex (i.e., including different types of effect and/or trend) the effect of autocorrelation becomes curvilinear and the $R^2$ variation diminishes for all indices.

*Effect of data pattern*

The exploration of data patterns' detection was carried out by constructing graphs combining the six procedures (PND, Cohen's $d_A$ and $d_{AB}$, Gorsuch's Trend analysis, White et al.'s $d$, Allison & Gorman's procedure) for computing the magnitude of effect with the six series lengths. In each of these 6 * 6 = 36 graphs we put data patterns in the abscissa and the effect size index ($R^2$ or percentage) in the ordinate, superimposing several autocorrelation levels. The ideal pattern of effects' detection would be represented by greater effect sizes for combined level and slope change, followed by second greater values for each of those effects separately and smaller values for data with no effect. A perfect index would not be affected by general trend not related to treatment's introduction. Therefore, greater discrepancy in $R^2$ or percentage between effects of interest and the remaining conditions meant better differentiation and indicated a more desirable performance.

The visual inspection carried out following those criteria suggests that the regression-based indices differentiate data patterns only for long and balanced series ($n_A = n_B = 10$ or 15), while also producing greater $R^2$. $d_A$ and $d_{AB}$ differentiate more than White et al.'s and Allison & Gorman's indices, being $d_{AB}$ the index that produces lower estimates of the magnitude of effect. PND proved to be the measurement that detected the most the differences between patterns even for short series ($n_A = n_B = 5$). A common problem of PND and the standardized mean differences is that they produce greater effect sizes in presence of trend (extraneous variable) than in presence of level change (intervention effect). As expected, complex patterns are associated with greater effect sizes for all indices.

As shown on Figure 3, Cohen's *d* are more sensitive to differential patterns. Nevertheless, the effect size values obtained through $d_A$ and $d_{AB}$ are smaller than the ones obtained via the regression-based procedures. Thus, the former indices have a lower probability to produce great effect sizes in absence of effects, a finding that becomes more evident in longer series. Figure 4 illustrates the higher differentiation between patterns accomplished by PND – the index that seemed to approximate better the ideal pattern described above. The figures show examples for $\varphi_1$ = 0.3, as it represents a level of serial dependence likely to be found in behavioral data (Parker, 2006), but the abovementioned tendencies were found for all $\varphi_1$ values simulated.

INSERT FIGURE 3 ABOUT HERE

INSERT FIGURE 4 ABOUT HERE

### *Series length effect*

Results' analysis revealed that incrementing series length leads to a higher differentiation between the data patterns. This, however, does not imply obtaining greater $R^2$. Actually we found that simple patterns (containing only one type of effect) produce higher estimations for $n_A = n_B = 5$ and $n_A = 5$, $n_B = 15$ than for $n_A = n_B = 10$ and 15. Consistent with the data simulation method, greater effect sizes we obtained for the (incremental) change in slope than for the (constant) change in level. As mentioned earlier, for the regression-based indices the values of $n_A$ and $n_B$ (and the relation between those) are relevant as it affects patterns distinction.

## Discussion

The purpose of the present study was to explore the performance of different effect size indices applied to data with known parameters. In applied settings it is frequent to have only few behavioral measurements which can be sequentially related. Therefore, the most useful indices to summarize the magnitude of the treatment effect will be the ones sensitive to effects in short data series, while being less affected by serial dependence. Out of the indices studied, the ones that performed better in the aforementioned terms were PND and standardized mean differences ($d_A$ and $d_{AB}$). Other advantages of these indices are calculus easiness and the fact that they are more widely known (especially, *d*) in comparison to regression-based procedures – a feature that might make them more attractive to applied researchers with lower degree of expertise in statistics. These indices differentiate better between the distinct data patterns and appear to have lower probability of false alarms in absence of treatment effect, but their results are distorted by trend. Hence, visual inspection can be used to detect trend and outliers prior to deciding whether the *d* and PND are appropriate effect size measures. A modification in the latter index will enable its application in cases when reduction rather than increment in the behavior of interest is expected. Recent proposals, related to the PND are the *Percentage of data points exceeding the mean* (Ma, 2006) and the *Percentage of all non-overlapping data* (Parker, Hagan-Burke, & Vannest, 2007) and their properties require further research.

It was surprising to find that the more sophisticated indices conceptualized for single-case designs (i.e., taking into account trend, level and slope change) performed worse than simpler and theoretically less appropriate strategies. Thus, future investigation is necessary to improve regression-based indices. Meanwhile, the use of simpler indices in N = 1 designs can be recommended whenever complementary information about trend is also taken into consideration. A possible source for additional information is visual analysis, which can enhance the choice of an appropriate effect size index and validate the results obtained by it (Parker, Cryer, & Byrns, 2006).

Among the limitations of the study we have to mention that only AB designs were studied due to their applicability in non-reversal behaviors. Nevertheless, the results presented here can be useful also for multiple-baseline designs for which there can be an effect size computed for each baseline (Busse, Kratochwill, & Elliott, 1995).

It has to be commented that the values of $\beta_1, \beta_2,$ and $\beta_3$ were not extracted from a previously published investigation due to the lack of indication in scientific literature. Apart from the $\beta$ values discussed, we also tried $\beta_2 = 0.6$ and $\beta_2 = 0.9$, varying the $\beta_1$ and $\beta_3$ values according to the formulae presented.

Very similar results were obtained and, as expected, all procedures showed greater discrimination between patterns (Figure 4 shows an example for one of the best performing indices). Nevertheless, future studies may continue exploring the optimal values of $\beta_1$, $\beta_2$, and $\beta_3$ for simulating different magnitudes of different data patterns. Another possible line of research is the application of the effect size indices to more-phased designs (e.g., ABAB) which are more suitable for controlling extraneous variables. In such a study it would be interesting to explore the variations in effect size as a function of how it was calculated: a) from phases $A_1$ and $B_1$; b) from phases $A_1$ and $B_2$; c) using the means of both A and both B phases; d) calculating an effect size for each change in phase.

# References

Allison, D. B., Faith, M. S., & Franklin, R. (1995). Antecedent exercise in the treatment of disruptive behavior: A review and meta-analysis. *Clinical Psychology: Science and Practice, 2*, 279-303.

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.

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education, 28,* 153-162.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meat-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.

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

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19,* 387-400.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

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

Faith, M. S., Allison, D. B., & Gorman, D. B. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement, 42*, 521-526.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, NJ: Erlbaum.

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

Greenwood, K. M., & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*, 107-118.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*, 291-304.

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

Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods, 39*, 343-349.

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767-786.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependence on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65,* 73-93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior modification, 30,* 598-617.

Mathur, S. R., Kavale, K. A., Quinn, M. M., Forness, S. R., & Rutherford, R. B., Jr. (1998). Social skills interventions with students with emotional and beahvioral problems: A quantitative synthesis of single-subject research. *Behavioral Disorders, 23,* 193-201,

Matyas, T. A. & Greenwood, K. M. (1990). Visual analysis for single-case time series: effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341-351.

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum Associates.

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). Useful effect size interpretations for single case research. *Behavior Therapy, 38,* 95-105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40,* 194-204.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32,* 879-883.

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

Scruggs, T. E., Mastropieri, M. A., Forness, S. R., & Kavale, K. A. (1988). Early language intervention: A quantitative synthesis of single-subject research. *The Journal of Special Education, 20,* 259-283.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavior Assessment, 10*, 243-251.

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.

Skiba, R. J., Casey, A., & Center, B. A. (1986). Nonaversive procedures in the classroom behavior problems. *The Journal of Special Education, 19,* 459-481.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9,* 113-124.

Suen, H. K., & Ary, D. (1987). Autocorrelation in behavior analysis: Myth or reality? *Behavioral Assessment, 9,* 150-130.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4,* 289-309.

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.

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

**Figure 1. Autocorrelation effect on different effect size measures.**

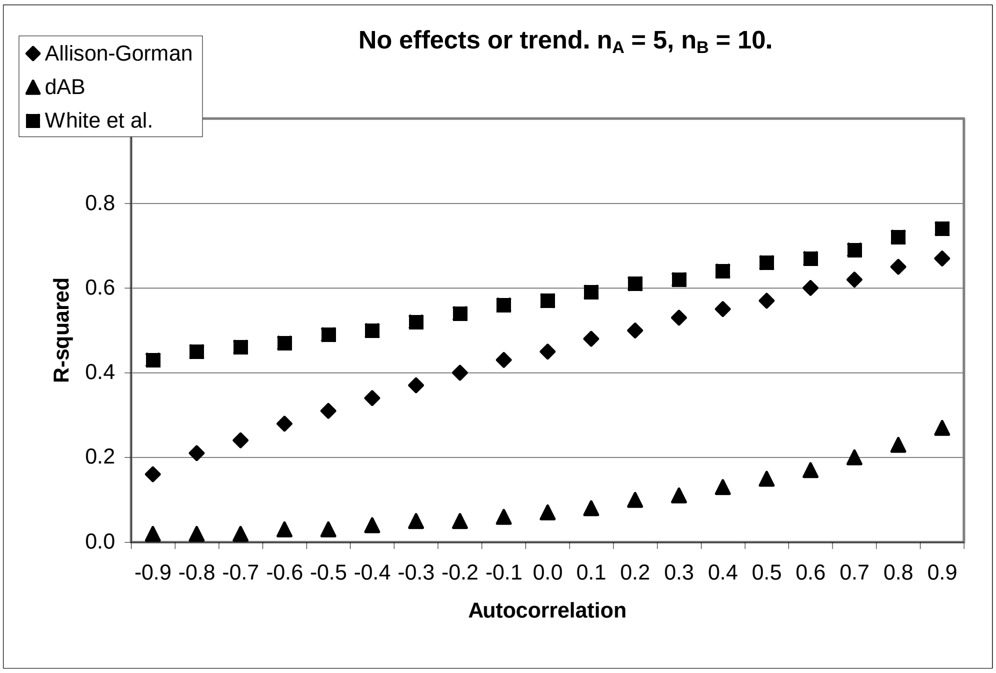


**Figure 2. Autocorrelation effect on the effect size calculated through the Percent of Nonoverlapping Data.**

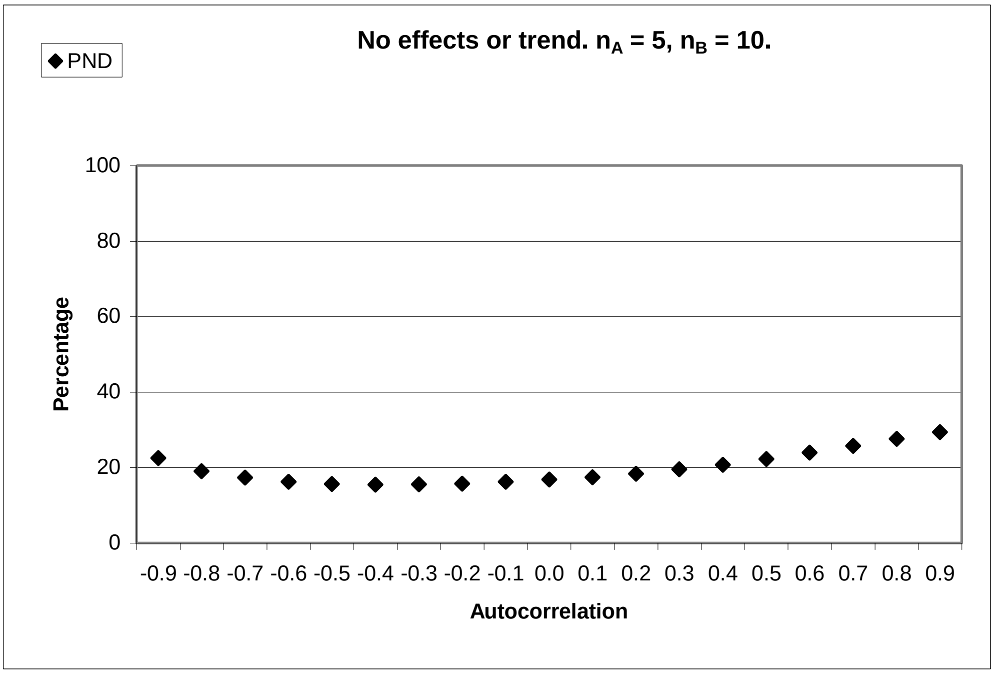


**Figure 3. Effect sizes calculated for different data patterns through two regression-based indices and one standardized mean difference index.**

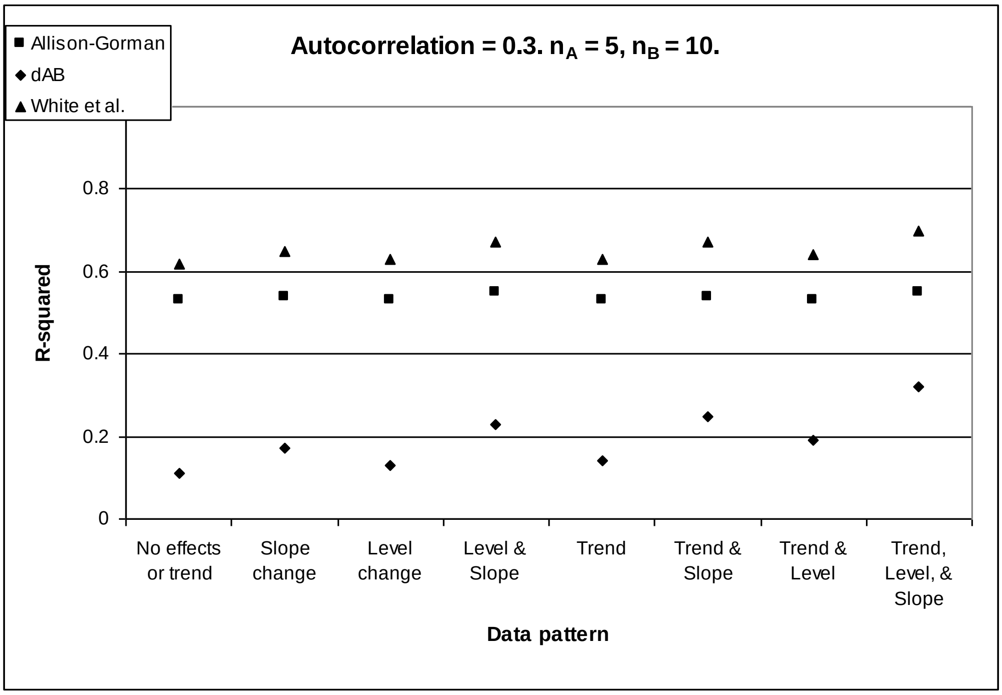
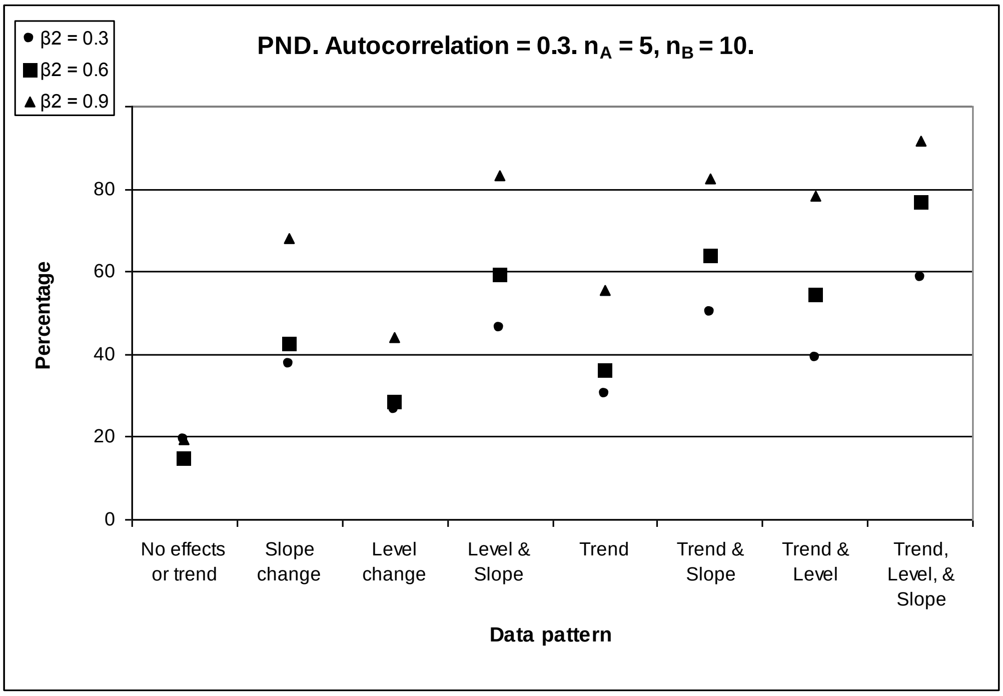


**Figure 4. Effect sizes calculated for different data patterns by means of the Percent of Nonoverlapping data.**


● β2 = 0.3
■ β2 = 0.6
▲ β2 = 0.9
PND. Autocorrelation = 0.3. $n_A$ = 5, $n_B$ = 10.
Percentage
No effects or trend
Slope change
Level change
Level & Slope
Trend
Trend & Slope
Trend & Level
Trend, Level, & Slope
Data pattern