Comparing “visual” effect size indices for single-case designs

Rumen Manolov, Antonio Solanas, and David Leiva

Department of Behavioral Sciences Methods,

Faculty of Psychology,

University of Barcelona

MAILING ADDRESS

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain. Phone number: +34933125844. Fax: +34934021359. Electronic mail may be sent to Rumen Manolov at rrumenov13@ub.edu.

AUTHORS' NOTE

This research was supported by the *Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa* of the *Generalitat de Catalunya*, the European Social Fund, the *Ministerio de Educación y Ciencia* grant SEJ2005-07310-C02-01/PSIC, and the *Generalitat de Catalunya* grant 2005SGR-00098.

RUNNING HEAD

N = 1 effect size indices

ABSTRACT

Effect size indices are indispensable for carrying out meta-analyses and can also be seen as an alternative for making decisions about the effectiveness of a treatment in an individual applied study. The desirable features of the procedures for quantifying the magnitude of intervention effect include educational/clinical meaningfulness, calculus easiness, insensitivity to autocorrelation, low false alarm and low miss rates.

Three effect size indices related to visual analysis are compared according to the aforementioned criteria. The comparison is made by means of data sets with known parameters: degree of serial dependence, presence or absence of general trend, changes in level and/or in slope. The percent of nonoverlapping data showed the highest discrimination between data sets with and without intervention effect. In cases when autocorrelation or trend is present, the percentage of data points exceeding the median may be a better option to quantify the effectiveness of a psychological treatment.

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

Single-case designs present problems for both data analysis of the specific study and quantitative integration of different studies. Replicating across subjects and settings in order to obtain evidence on the strength of the intervention is useful only when there are summary measures available to be used in meta-analyses.

The difficulties in single-case designs analysis are related to the scarce number of observations usually available (Huitema, 1985) and to the serial dependence between the measurements obtained from the same experimental unit (Busk & Marascuilo, 1988; Matyas & Greenwood, 1991; 1997; Parker, 2006). Whether being statistically significant or not, autocorrelation has been alleged to affect the analytical techniques employed (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Scientific evidence points out that serial dependence alters the performance of procedures as diverse as ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983), the split-middle method (Crosbie, 1987) and randomization tests (Gorman & Allison, 1997; Sierra, Solanas, & Quera, 2005). On the other hand, for determining the effectiveness of a treatment in an individual study it is not sufficient to obtain a *p*-value, due to the disadvantages of this indicator (Cohen 1990; 1994; Kirk, 1996; Rosnow & Rosenthal, 1989; Wilkinson & The Task Force on Statistical Inference, 1999). Clinical, educational and social researchers need more meaningful information than the one provided by the statistical significance. Visual analysis, as an alternative, is more subjective and does not allow quantification. Moreover, it has been found to be distorted

by the presence of serial dependence (Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990). An objective measurement that can be used to quantify the relationship between the treatment and the behavior of interest is effect size.

In contrast with *p*-values, effect size indices are useful for documenting results for posterior meta-analysis and power analysis (Parker & Hagan-Burke, 2007b). Among the advantages of effect size, the following have been stated: a) it is not systematically affected by sample size (Parker & Brossart, 2003); b) it uses on the strength of association between the independent and the dependent variables, instead of centering on the null hypothesis (Kromrey & Foster-Johnson, 1996); c) it allows treatments' comparison (Parker & Hagan-Burke, 2007b); and d) it is possible to construct confidence intervals about the effect size (Kirk, 1996).

The most widely known effect size indices based on standardized mean differences (e.g., Cohen's *d*; Hedges' *g*; Glass' Δ) and measurements of association (e.g., η^2; ω^2; R^2) were not developed for single-case designs but rather for designs involving groups' comparison and, thus, focus only on the average levels of behavior in the different conditions. Nonetheless, there are also procedures conceptualized for N = 1 designs – some of them based on regression analysis and others closely related to visual analysis. It is possible to convert some effect size indices into others (Friedman, 1982), allowing the comparison between meta-analyses using different measures. The bibliographic search we performed suggests that visually-based indices are

applied more often (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) in meta-analyses. This could be due to the advantages of visual indices, such as calculus easiness and increased interpretability from clinical and educational perspective.

Regression-based effect size indices

The regression-based procedures incorporate predictor variables in order to model changes in level and in slope and also try to control for extraneous variables such as trends. The following procedures are some of the most studied ones in scientific literature:

1) Gorsuch's (1983) trend analysis includes time as covariate and eliminates its influence prior to testing for change in level.

2) White, Rusch, Kazdin, and Hartmann's (1989) *d*, taking into consideration the correction presented in Faith, Allison, and Gorman (1997), compares two predicted values – the last treatment phase point according to baseline phase regression equation with the last treatment phase point as predicted by the treatment phase regression equation. The model also takes into account the possible relation between time and the measured behavior.

3) Center, Skiba, and Casey's (1985-1986) model, in contrast with the abovementioned procedures, can account for both changes in level and slope,

while controlling for the presence of trend. Among the limitations of this procedure have been stated the attainment of more than one magnitude of effect index and the impossibility to obtain a negative *d*.

4) Allison and Gorman's (1993) model pretends to improve the previous technique, estimating trend solely from the baseline phase and allowing the correspondence between the type of treatment effect (i.e., reducing or increasing the behavior of interest) and the sign of the effect size index (negative or positive, respectively). A shortcoming of the model is the possible effect size overestimation.

Commons drawbacks of the regression-based procedures are the parametric assumptions, while there is also evidence that despite of their conceptual appropriateness those models do not perform as well as simpler indices (Manolov & Solanas, 2008).

Visual effect size indices

These effect size indices are based on a criterion employed in visual analysis in order to decide the effectiveness of a treatment – the amount of overlap between the data points pertaining to baseline and treatment phases. Their attractiveness to applied researchers is related to calculation easiness and to the fact that visual inspection is still the most commonly applied single-case data analysis technique (Parker, Cryer, & Byrns, 2006). Some of the procedures proposed for using in psychological studies are:

1) Scruggs, Mastropieri, and Casto's (1987) percent of nonoverlapping data (hereinafter, PND). PND is based on the proportion of treatment phase measurements greater than the highest baseline phase data point. It has been criticized for ignoring all phase A data points except for one, a reason for which the following two indices were proposed.

2) Ma's (2006) percentage of data points exceeding the median (hereinafter, PEM). PEM was proposed to correct some of the potential drawbacks of PND, like the sensitivity to floor or ceiling effects, while maintaining its advantages. As its name suggests, this index computes the percentage of treatment measurements greater than the baseline phase median.

3) Parker, Hagan-Burke, and Vannest's (2007) percentage of all non-overlapping data (hereinafter, PAND). PAND was introduced as an alternative to PND for larger data sets. It takes into account all data points and counts the minimum number of measurements that need to be removed in order to obtain series with no overlap. The ratio between the remaining data points and series' length is the basis of the index. The authors also suggest that the index can be converted into a *Phi* effect size index or an improvement rate difference.

The objective of the present study was to extend the scientific literature (e.g., Parker & Hagan-Burke, 2007a) assessing the performance of the three measures of effect sizes for AB designs in presence of different degrees of autocorrelation. We aimed to explore which index discriminates better between the distinct data patterns, while an additional purpose was to evaluate

the influence of series' length, following Campbell's (2004) suggestions. As the estimation and hypothesis testing of serial dependence from real data can 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 (Manolov & Solanas, 2008; Parker & Brossart, 2003).

Method

Design selection

The study focused on AB designs with several series' lengths (N) and phase lengths (n_A and n_B), short enough to be feasible in applied settings where the temporal cost has to be taken into consideration. We chose the following values in order to cover a range of possible "short series":

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

For each series' length we generated data sets with different patterns, defined by the presence or absence of general trend, change in level and/or in slope. The statistical model used was suggested by Huitema and 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;

β_0: intercept;

β_1: coefficient associated with general trend;

β_2: coefficient associated with level change;

β_3: coefficient associated with slope change;

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

D_t: dummy variable for level change. For phase A it was set to 0 and for phase B to 1;

SC_t: value of the slope change variable, computed as $[T_t - (n_A + 1)] * D_t$, so that it is equal to 0 for phase A, and takes values from 0 to $(n_B - 1)$ for phase B;

ε_t: error term;

The error term (ε_t) was generated following a first-order autoregressive model: $\varepsilon_t = \varphi_1 * \varepsilon_{t-1} + u_t$. The values of serial dependence (φ_1) ranged from –.9 to .9 in steps of .1. The u_t term represents white noise at moment t generated following N (0, 1) and $\varepsilon_1 = u_1$.

The value of the intercept parameter β_0 was set to zero as it does not affect effect size calculation. In order to ensure the adequacy of the comparison

between experimental conditions, we chose the values of β_1, β_2, and β_3 so that they produce comparable mean differences between the two phases. We chose to set first the β_2 parameter, as the level change is maintained constant throughout the whole intervention phase. Afterwards, we set the values of β_1 and β_3 leading to the same difference $\bar{y}_B - \bar{y}_A$. Those steps were initially carried out for the shortest series (i.e., $n_A = n_B = 5$) in order to explore if longer series imply better discrimination of data patterns. We tested several values for β_2 (from .1 to .6 in steps of .1) for all experimental conditions seeking its most appropriate value. We found that for $\beta_2 = .1$ the values of PND were all too low, while for $\beta_2 = .6$ PEM was close to reaching its maximum value. To avoid the floor and ceiling effects (see Figure 1), which make impossible patterns discrimination, we decided to set β_2 to .3.

INSERT FIGURE 1 ABOUT HERE

The use of $\beta_2 \neq 0$ implies that $\bar{y}_B - \bar{y}_A = \beta_2$ if the other parameters are set to zero. The value of β_3 that leads to the same mean difference can be found through the following expression:

$\beta_3 = \frac{\beta_2}{\frac{n_B - 1}{2}}$, which for $\beta_2 = .3$ leads to $\beta_3 = \frac{.3}{\frac{5-1}{2}} = \frac{.6}{4} = .15$,

while the appropriate value of β_1 is obtained as:

$\beta_1 = \frac{\beta_2}{\frac{n_A + n_B}{2}}$, which for $\beta_2 = .3$ leads to $\beta_1 = \frac{.3}{\frac{5+5}{2}} = \frac{.6}{10} = .06$.

We could verify that the β_1 and β_3 values are appropriate for producing mean differences equal to the value of β_2 even for the most extreme levels of serial dependence (−.9 and .9), whenever $n_A = n_B = 5$. In total there were eight data patterns studied, defined by the presence and combination of trend, level change, and slope change (i.e., β_1, β_2, and β_3 being equal to or different from zero).

Finally, in order to guarantee suitable simulated data, the 50 values 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

Prior to presenting in detail the steps needed to compute the three effect size indices included in the present study, an example of a fictitious data set is presented. Consider a psychological study applying the Parent Child Interaction Therapy (for an in-depth description see Borrego, Anhalt, Terao, Vargas, & Urquiza, 2006) in which the number of praises a parent directs to a child is registered five days prior to treatment introduction and five days during intervention. The data gathered using the AB design structure (4, 5, 3, 6, and 3 praises during baseline and 7, 5, 8, 9, and 7 praises during treatment phase) can be represented graphically as shown on Figure 2. In following

section, each of the procedures is applied to the data set presented in order to illustrate their calculus.

INSERT FIGURE 2 ABOUT HERE

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

Percent of nonoverlapping data:

1) Identify the highest measurement in phase A. In the example it is 6 praises corresponding to baseline day 4.
2) Calculate the number of phase B data points that exceed the value identified in the previous step. The measurements corresponding to days 6, 7, 9, and 10 are greater than 6, so there are 4 values exceeding phase A's highest value.
3) Divide the value obtained in step 2 by the number of observations in phase B. The number of phase B observations is 5 and the result of the division is 4/5 = 0.8.
4) Multiply the value obtained in step 3 by 100 in order to convert the proportion into a percentage. The percentage obtained for the example is 0.8*100 = 80%.

Percentage of all non-overlapping data:

1) Identify the highest measurement in phase A. As obtained above this value is 6.

2) Calculate the minimal number of data points to be eliminated in order to have no inter-phase overlap. If the measurement corresponding to day 7 (i.e., 5 praises) is eliminated, then phase A and phase B would not overlap – all phase B data points would be greater than the phase A measurements.
3) Divide the value obtained in step 2 by the total number of observations. A single value to be eliminated means that the correct division is 1/10 = 0.1.
4) Multiply the value obtained in step 3 by 100. The value obtained is 0.1*100 = 10%.
5) Subtract the value obtained in step 4 from 100. The percentage of all data non-overlapping data is equal to 100 – 10 = 90%.

Percentage of data points exceeding the median:

1) Calculate the median of phase A. In the example, the sorted baseline measurements are 3, 3, 4, 5, and 6 and, therefore, the phase A median is equal to 4.
2) Calculate the number of phase B data points that exceed the value identified in the previous step. All data points from the treatment phase are greater than 4, so the value obtained is 5 (equal to n_B).
3) Divide the value obtained in step 2 by the number of observations in phase B. The division to be made is 5/5 = 1.

4) Multiply the value obtained in step 3 by 100 in order to convert the proportion into a percentage. In the example presented, the percentage of data points exceeding the median obtained is, thus, 1*100 = 100%.

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 (β_1, β_2, and β_3) parameters for data generation, leading to 8 different data patterns – autoregressive model (i.e., no effect or trend); 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 15.
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 ε_t using the equation $\varepsilon_t = \varphi_1 * \varepsilon_{t-1}$.
9) Obtain the time array $T_t = 1, 2, \ldots, N$.

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 and McKean's (2007) expression: $SC_t = [T_t - (n_A + 1)]*D_t$ used for data generation.

11) 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 PND.

14) Calculate PAND.

15) Calculate PEM.

16) Average the obtained percentages from the 100,000 replications of each experimental condition.

Results

This section is organized according to the objectives of the study: to explore the effect of autocorrelation, to compare data patterns discrimination, and to assess the importance of series' length.

Autocorrelation effect

In order to quantify the degree to which autocorrelation introduces distortion in the effect size estimates, we divided the estimates obtained for $\varphi_1 \neq 0$ by the one obtained for $\varphi_1 = 0$. We performed those calculi for the case of no effect

or trend simulated to avoid confounding variables. If the ratio obtained is equal to 1, then there is no influence of serial dependence. Ratios lower than 1 imply an underestimation of the effect size associated with autocorrelation, while values greater than 1 entail overestimation. As Table 1 shows, PEM yields practically the same values regardless of the degree of serial dependence. For PND and PAND greater negative or positive autocorrelation is generally associated with higher effect size estimates, being PND the more affected of the two indices. Figure 2 shows and example of those findings.

INSERT TABLE 1 ABOUT HERE

INSERT FIGURE 3 ABOUT HERE

When there was treatment effect simulated in data, PEM proved to be sensitive to the presence of autocorrelation – positive as well as negative serial dependence leads to lower effect size estimates (see Figure 3 for an example). For PND and PAND, the type of relationship between autocorrelation and effect size depends on the type of effect in data. When the intervention involves a level change, positive and negative φ_1 overestimate effect size. When the treatment effect is expressed as slope change, it would be underestimated if PND or PAND are used. Figure 4 is an illustration of these tendencies.

INSERT FIGURES 4 AND 5 ABOUT HERE

Data pattern discrimination

The comparison of data patterns discrimination was carried out by constructing graphs combining the three procedures for computing the magnitude of effect with the six series' lengths. In each of these 3 * 6 = 18 graphs we put data patterns in the abscissa and the effect size index (i.e., percentage) in the ordinate, superimposing several autocorrelation levels.

We consider that an effect size index should detect (i.e., yield highest effect size estimates) powerful treatments, like the ones represented by changes in slope and in level in the same direction. The indices would also have to respond with high estimates to the occasions when either a change in level or a change in slope is present. On the other hand, when the intervention is not effective the effect size index ought to yield low (ideally zero) percentages. Additionally, a perfect index would not be sensitive to a general trend, which has no relation to the introduction of a psychological treatment.

The visual inspection carried out following those criteria suggests that PND and PEM approximate the ideal discrimination pattern. Nonetheless, there is one relevant discrepancy between those two indices due to the essence of their calculus – PND yields smaller effect size estimates than PEM. PAND seems to be more deficient, as it yields more similar estimates for data sets with and without treatment effects. An example of those findings can be seen in Figure 5, which is constructed for φ_1 = .3, as it represents a level of serial dependence likely to be found in behavioral data (Parker, 2006), although the

abovementioned tendencies are common to all φ_1 values studied. All of the indices tested share a common drawback – they are affected by the presence of trend in data which leads to overestimating effect size. As expected, complex patterns are associated with greater effect size estimates for all indices.

INSERT FIGURE 6 ABOUT HERE

Complementing the analyses performed, we divided the effect size estimates for series with effect and/or trend present by the estimate for data with no effect or trend simulated. These calculi were carried out for each of the three indices and for all series' lengths. Ratios equal to 1 suggest that there are the same estimates obtained in presence and in absence of effect. Values greater than 1 imply that the effect or the extraneous variable are associated with greater effect size estimates than white noise data. As Table 2 shows, PND is the procedure that differentiates the most between presence and absence of intervention effect. However, it is also the procedure most affected by trend. PAND distinguishes less between data patterns, except for data series with $n_A = 5$ and $n_B = 15$ where its performance is practically equivalent to PEM's.

INSERT TABLE 2 ABOUT HERE

Series' length effect

In order to explore the variation of the performance of the indices as one of the phases (or both) becomes longer, we divided the effect size estimates obtained for the longer designs with the ones obtained for the shortest one ($n_A = n_B = 5$). Ratios equal to 1 suggest that phase length does not influence the performance of the procedures. Values greater or smaller than 1 imply higher or lower effect size estimates, respectively, in comparison to 10-measurements data sets. According to Table 3, increasing series' length leads to a better differentiation between the data patterns. As the example in Figure 6 shows the improvement is expressed basically as lower false alarm rates (i.e., lower percentages for the case of absence of treatment effect) and as higher sensitivity to synergic slope and level changes. Those results highlight the importance of having more measurements of the experimental unit in order to obtain a more precise image of the evolution of its behavior. In accordance with the data simulation method followed, in longer series changes in slope yielded higher effect size estimates than changes in level.

INSERT TABLE 3 ABOUT HERE

INSERT FIGURE 7 ABOUT HERE

The performance of PAND improves for designs with unbalanced phase lengths. As Figure 7 illustrates for such designs the distinction between data patterns is more pronounced, implying lower effect size estimates for white noise and trend. On the contrary, for PND the presence of trend is more

problematic for designs with unequal phase lengths. PEM is the procedure less affected by the amount of data points in the series.

INSERT FIGURE 8 ABOUT HERE

Discussion

In the current investigation we pretended to continue the search of the most appropriate procedure for quantifying treatment effectiveness and summarizing results from single-case designs. The performance of the effect size indices was tested by means of data patterns generated to represent the likely features of real data (i.e., few observations per phase, serially dependent measurements). Among the desirable features those indices can be stated: a) to detect changes in behavior due to the introduction of an intervention – low miss (Type II error) rates; and b) to produce low, ideally null, effect size estimates in absence of treatment effect – low false alarm (Type I error) rates; c) to be insensitive to extraneous variables such as general trend; and d) to remain unaffected by autocorrelation.

Taking the first two criteria into consideration simultaneously we can point to PND as the best performer as it produces lowest effect size estimates in presence of solely white noise. Moreover, among the three procedures tested, it presents the highest relative differentiation between effective and ineffective interventions. PEM also shows a good patterns' discrimination, being more

sensitive but less specific than PND. PAND is the index that performs less satisfactorily in the cases when baseline and treatment phases have approximately the same number of observations. A positive characteristic of all three indices studied is the discrimination between data patterns even when series consist of only ten data points.

As regards autocorrelation, PEM is the less affected procedure in absence of effect and is conservatively biased by both positive and negative serial dependence in presence of treatment effect. Applied researchers should keep in mind that both overestimation and underestimation of an existing treatment effect are possible when PND and PAND are used, depending on the degree of autocorrelation and on the type of effect (change in slope or in level). Out of those two indices PND is the one whose effect size estimates are more distorted by serial dependence.

A shortcoming of the indices is the finding of the distorting impact of trend in data, which makes necessary the visual inspection prior to applying any of the three procedures. PAND was the least affected index, while PND was the most affected one.

In conclusion, what recommendation can be given to applied researchers? To begin with, they ought to keep in mind what each index represents in order to interpret it correctly. In this sense, we consider that the meaning of PND and PEM is more straightforward that the information given by PAND. In terms of computational accessibility, all three indices can easily be calculated, especially PND. We have to advert that whenever the intervention is supposed

to reduce rather than to enhance the behavior measured, the manner of computation of the indices can be adjusted to the needs of the applied researcher. A potential advantage of PAND is the possibility to derive from it a conventional effect size index, like Pearson's *Phi* (Parker et al., 2007). Nonetheless, mathematical-statistical calculations beyond the computation of the percentage itself may make the index less attractive to applied researchers. Applied researchers can be advised to use PND in data sets with no autocorrelation or trend, as it is the procedure that best distinguishes between presence and absence of intervention effect. When there is a high outlier in the baseline phase and the objective of the intervention is to increase the behavior of interest, the use of PND cannot be advised as it would lead to an underestimation of the treatment effect. In cases when the behavioral measurements present general trend or are likely to be sequentially related, PEM ought to be the effect size index chosen. PAND approximates PEM's performance only when the baseline phase is considerably shorter than the treatment phase.

In any case, professionals should not follow the same criteria for labeling the treatment as "effective" when using different procedures (e.g., 70%-90% "effective", 50%-70% "questionable", in Scruggs and Mastropieri, 1998). This is due to the fact that as some of the indices (PEM and PAND) yield systematically higher effect size estimates than others (PND). Whatever index is utilized, visual inspection should not be replaced as a source of supplementary information (Parker et al., 2006).

As regards meta-analysis of single-case data, applied psychologists ought to be cautious when integrating information from studies using different number of measurement times, since these may imply different levels of affection by autocorrelation and general trend. That is, the effect size estimates obtained from studies with a specific N may not have the same precision and the same insensitivity to extraneous variables as the estimates obtained for other series and/or phase lengths. This difficulty is, however, not only applicable to effect size procedures based on visual analysis, but also to the ones based on regression or standardized mean difference (Manolov & Solanas, 2008).

A limitation of the present investigation consists in the fact that only two-phase designs were studied. However, as Busse, Kratochwill, and Elliott (1995) claim, the AB designs' results can also be useful for multiple-baseline designs.

Future research may center on calibrating the data generation procedure with the most appropriate values (i.e., β_1, β_2, and β_3) for simulating treatment effects in order to improve real data modeling. In addition, it is necessary to obtain evidence on the performance of the effect size indices in designs consisting of more than two phases.

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.

Borrego, J., Jr., Anhalt, K., Terao, S. Y., Vargas, E. C., & Urquiza, A. J. (2006). Parent-child interaction therapy with a Spanish-speaking family. *Cognitive and Behavioral Practice, 13*, 121-133.

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). Meta-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. (1997). 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.

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. (1997). 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: Lawrence 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.

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

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 behavioral 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. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137-157.

Matyas, T. A., & Greenwood, K. M. (1997). 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.

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. (2007a). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

Parker, R. I., & Hagan-Burke, S. (2007b). 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., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, 8, 24-33.

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.

Tables

Table 1. Distortion due to autocorrelation when no trend or effect is present in data – the values represent the ratio $\varphi_1 \neq 0 / \varphi_1 = 0$.

	Effect size	**Series' length**					
φ	**indices**	**5+5**	**5+10**	**7+8**	**5+15**	**10+10**	**15+15**
-.9	PND	1.32	1.36	1.46	1.38	1.60	1.78
	PAND	1.05	1.09	1.06	1.13	1.05	1.05
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
-.6	PND	.97	.97	.99	.97	1.02	1.06
	PAND	1.00	.99	1.00	.99	1.00	1.00
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
-.3	PND	.93	.93	.93	.93	.94	.97
	PAND	.99	.98	.99	.98	1.00	1.00
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
.3	PND	1.16	1.17	1.17	1.16	1.17	1.18
	PAND	1.02	1.04	1.02	1.05	1.01	1.01
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
.6	PND	1.38	1.44	1.50	1.47	1.58	1.64
	PAND	1.05	1.11	1.06	1.16	1.05	1.04
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
.9	PND	1.61	1.76	1.94	1.83	2.27	2.86
	PAND	1.09	1.19	1.12	1.28	1.11	1.11
	PEM	1.00	1.00	1.00	1.00	.99	1.00

Table 2. Detection of data patterns in comparison to the case of no effect or trend simulated in independent series.

	Effect size	Series' length					
Data pattern	**indices**	**5+5**	**5+10**	**7+8**	**5+15**	**10+10**	**15+15**
Slope change	PND	1.45	2.12	2.01	2.82	2.67	4.89
	PAND	1.06	1.28	1.13	1.61	1.14	1.23
	PEM	1.21	1.42	1.35	1.57	1.44	1.60
Level change	PND	1.42	1.43	1.49	1.43	1.57	1.66
	PAND	1.06	1.11	1.06	1.14	1.05	1.04
	PEM	1.21	1.21	1.21	1.21	1.22	1.23
Level & slope	PND	1.95	2.66	2.69	3.35	3.60	6.23
Change	PAND	1.14	1.42	1.21	1.78	1.22	1.31
	PEM	1.40	1.58	1.52	1.70	1.60	1.72
Trend	PND	1.43	1.67	1.77	1.94	2.27	3.59
	PAND	1.06	1.17	1.10	1.31	1.11	1.15
	PEM	1.21	1.30	1.31	1.39	1.42	1.60
Trend &	PND	1.95	2.93	3.07	3.75	4.58	8.71
slope change	PAND	1.14	1.48	1.26	1.92	1.30	1.45
	PEM	1.39	1.61	1.58	1.74	1.70	1.84
Trend &	PND	1.92	2.21	2.46	2.51	3.17	4.98
level change	PAND	1.13	1.30	1.18	1.50	1.18	1.23
	PEM	1.40	1.49	1.50	1.56	1.59	1.73
Trend, level, &	PND	2.50	3.47	3.79	4.17	5.56	9.97
slope change	PAND	1.21	1.62	1.35	2.05	1.38	1.52
	PEM	1.56	1.73	1.71	1.82	1.80	1.90

Table 3. Influence of series' length on pattern detection for independent series - comparison to $n_A = n_B = 5$.

Series' length	Effect size indices	Data pattern				
		No effect or trend	**Slope change**	**Level change**	**Level & slope change**	**Trend**
5+10	PND	1.00	1.47	1.00	1.37	1.17
	PAND	.76	.92	.79	.95	.84
	PEM	1.00	1.18	1.00	1.13	1.08
7+8	PND	.75	1.04	.78	1.04	.93
	PAND	.91	.97	.91	.98	.94
	PEM	1.00	1.12	1.00	1.09	1.09
5+15	PND	1.00	1.95	1.00	1.72	1.36
	PAND	.64	.97	.69	1.01	.80
	PEM	1.00	1.31	1.00	1.22	1.15
10+10	PND	.55	1.01	.60	1.01	.87
	PAND	.94	1.00	.92	1.00	.97
	PEM	1.00	1.20	1.01	1.15	1.18
15+15	PND	.37	1.26	.43	1.19	.94
	PAND	.91	1.05	.89	1.05	.99
	PEM	1.00	1.33	1.01	1.23	1.32

Figures

Figure 1. Influence of the simulation parameters β on the effect size indices.

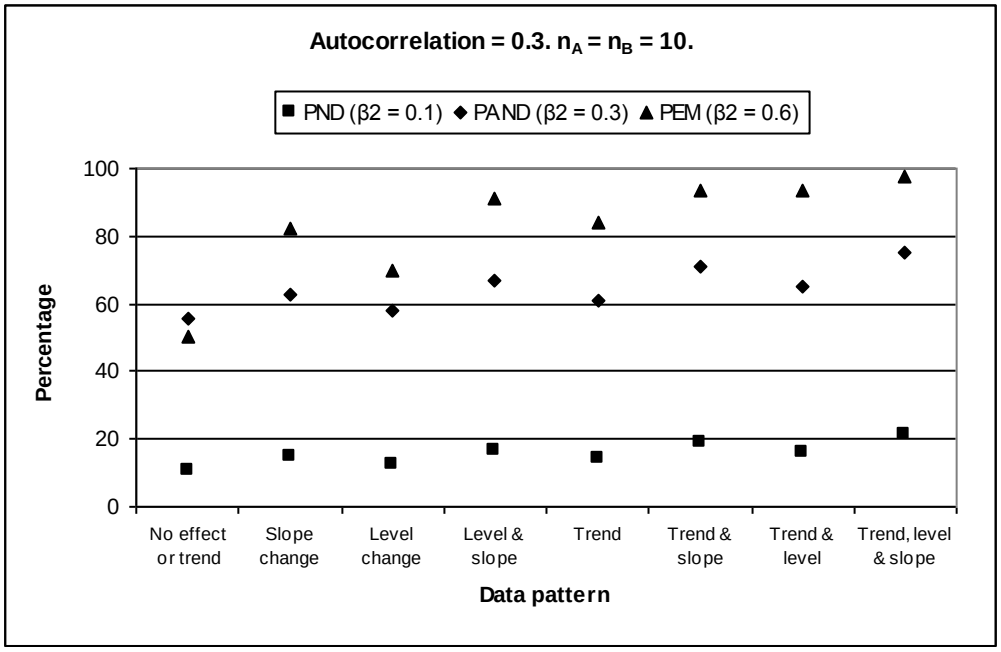

Figure 2. A fictitious example of an AB data series with $n_A = n_B = 5$.

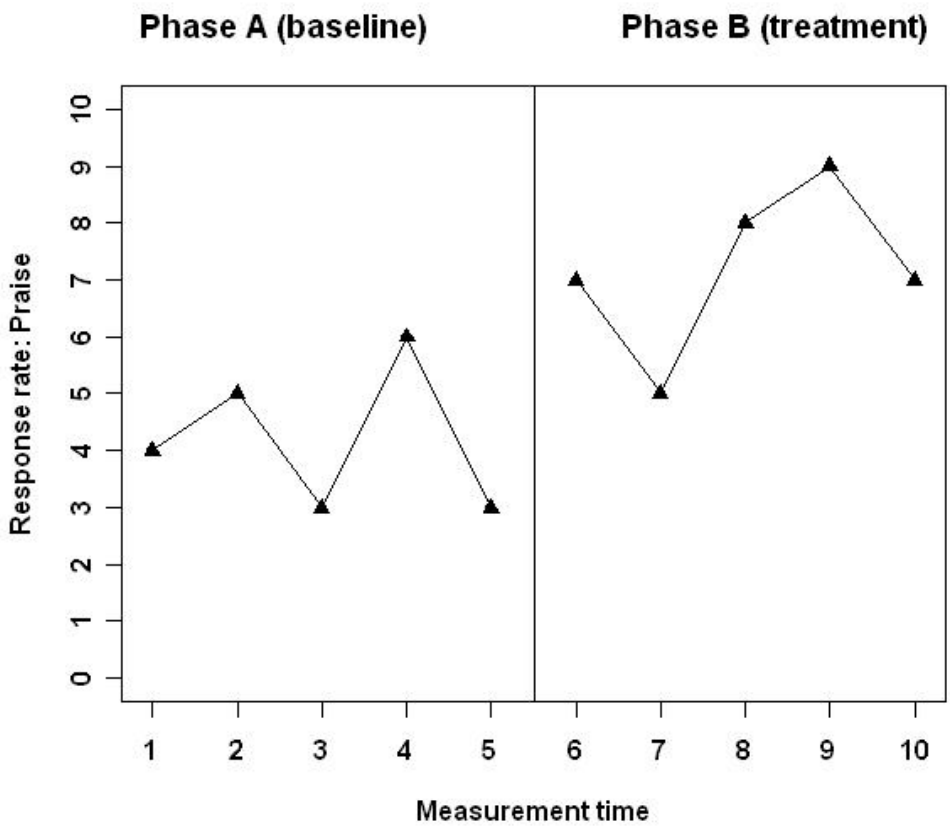

Figure 3. Autocorrelation effect on the effect size indices when no effect or trend are present in data.

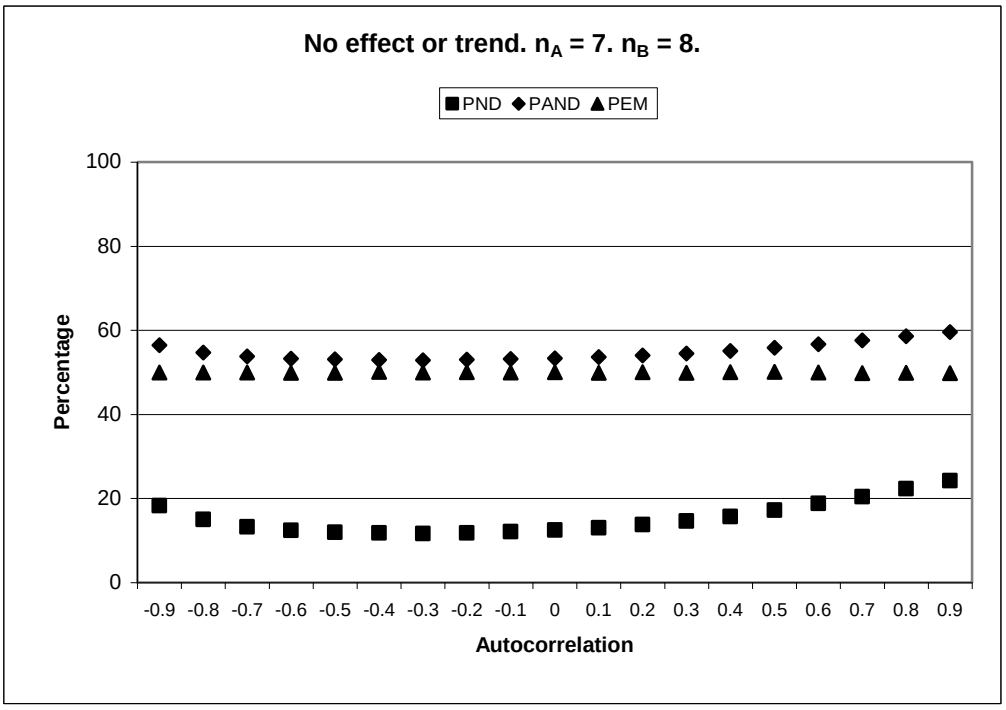

Figure 4. Autocorrelation effect on PEM when treatment effects are present in data.

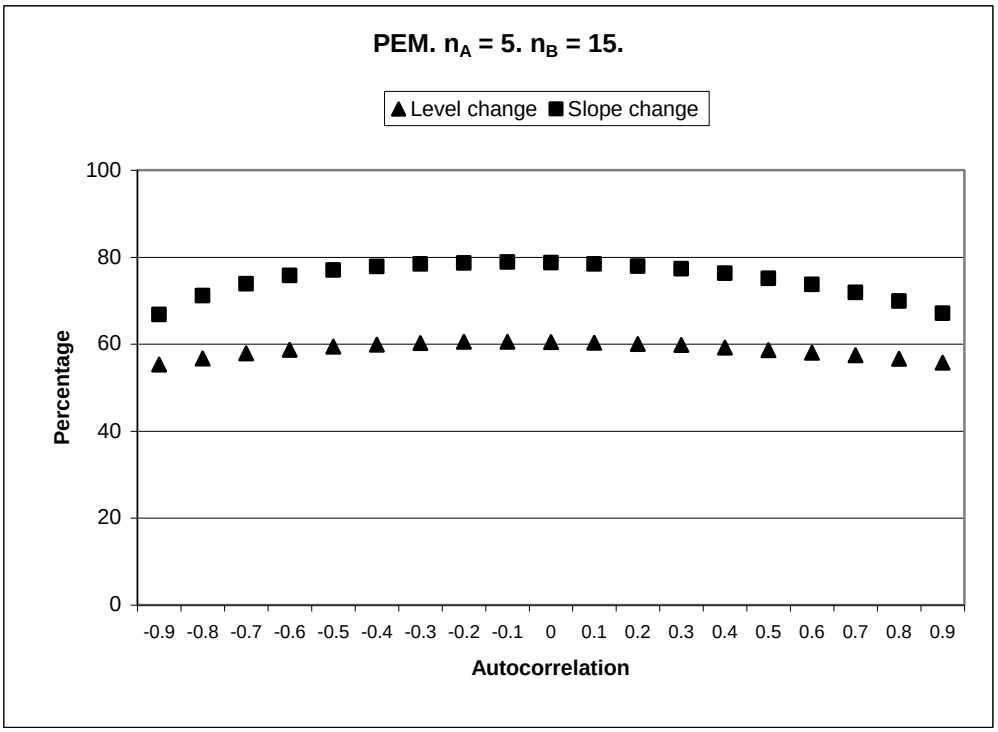

Figure 5. Autocorrelation effect on PND when treatment effects are present in data.

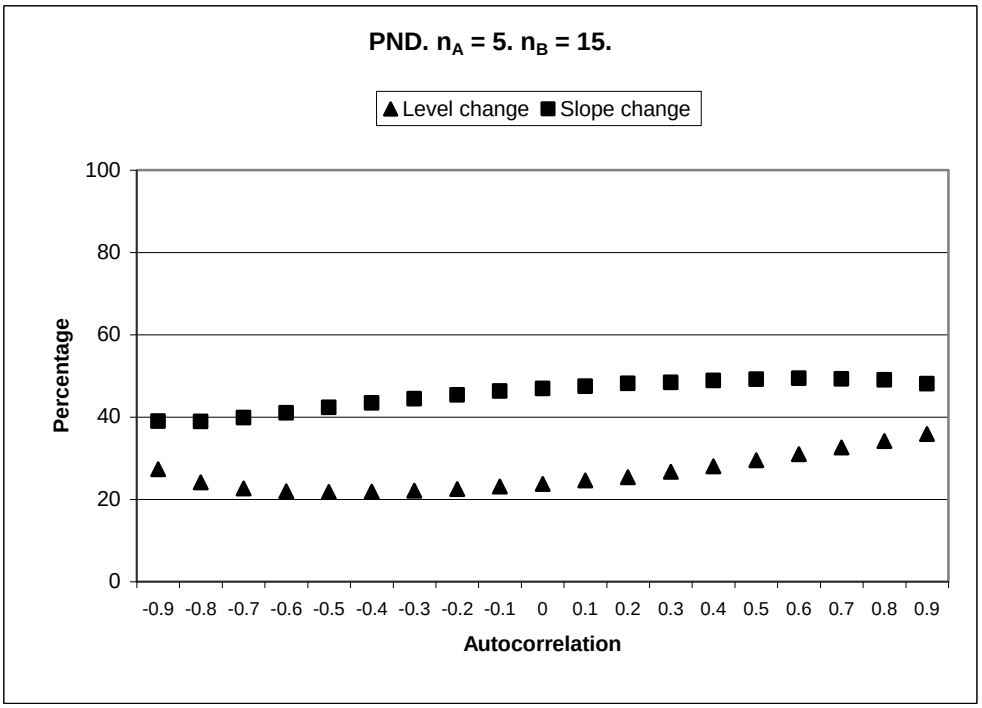

Figure 6. Effect sizes calculated for different data patterns and moderate positive serial dependence in a design with equal phase lengths.

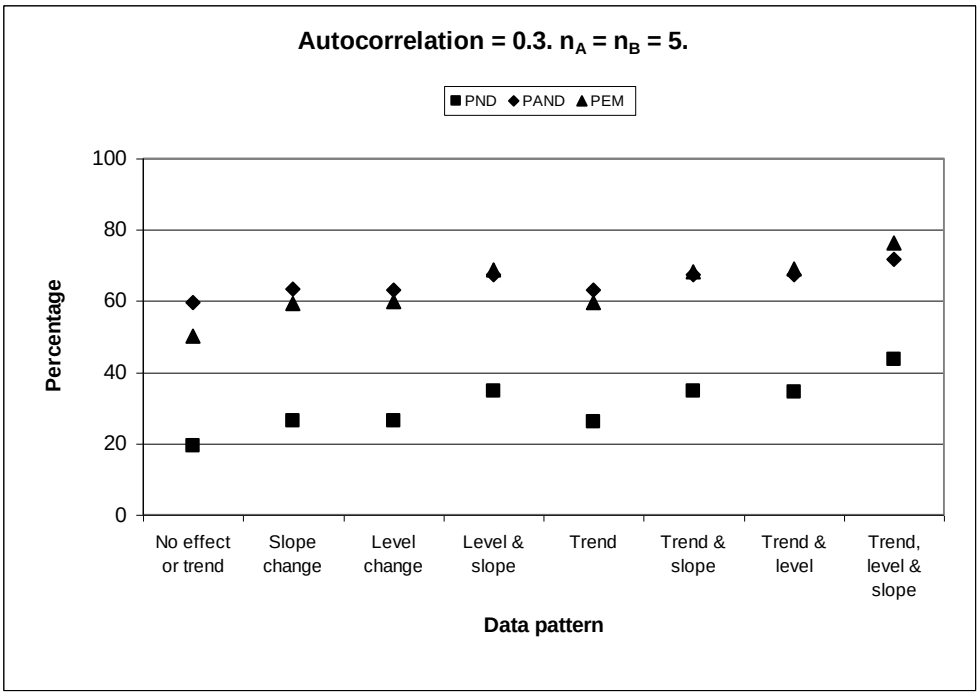

Figure 7. Influence of series' length on PND.

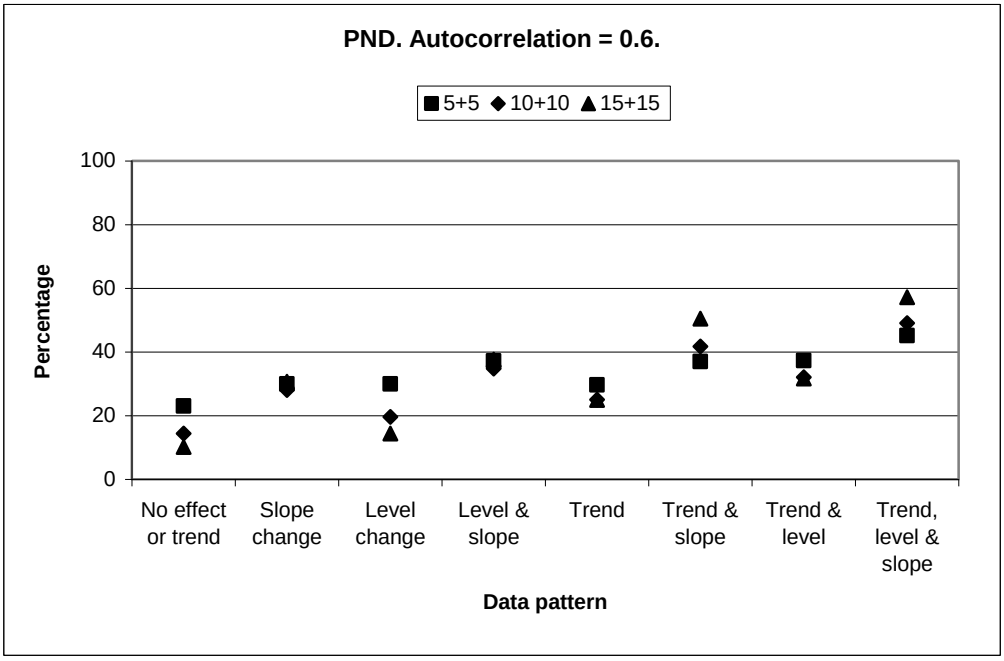

Figure 8. Influence of phase length on PAND.