SECCIÓN METODOLÓGICA

Psicológica (2009), 30, 137-154.

Problems of the randomization test for AB designs

Rumen Manolov* and Antonio Solanas

University of Barcelona

N = 1 designs imply repeated registrations of the behaviour of the same experimental unit and the measurements obtained are often few due to time limitations, while they are also likely to be sequentially dependent. The analytical techniques needed to enhance statistical and clinical decision making have to deal with these problems. Different procedures for analysing data from single-case AB designs are discussed, presenting their main features and revising the results reported by previous studies. Randomization tests represent one of the statistical methods that seemed to perform well in terms of controlling false alarm rates. In the experimental part of the study a new simulation approach is used to test the performance of randomization tests and the results suggest that the technique is not always robust against the violation of the independence assumption. Moreover, sensitivity proved to be generally unacceptably low for series lengths equal to 30 and 40. Considering the evidence available, there does not seem to be an optimal technique for single-case data analysis.

In psychological research there seem to be basically two ways of carrying out a study. The first one involves comparing groups before and after a treatment has been administered to one of them and is usually referred to as "group designs". The second implies repeated measurements of the same individual or group taken as a unity under different conditions and it is generally labelled as "single-case designs". N = 1 designs have

* 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. The authors would like to thank the anonymous reviewers for their useful comments and suggestions, which contributed to improving the manuscript. 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. E-mail: rrumenov13@ub.edu

certain advantages as they allow studying the evolution of a unit in time and they permit addressing idiosyncrasy. Moreover, single-case designs are more feasible when the population of interest is small or disperse and groups cannot be easily formed.

There are several ways of conducting a single-case study and the most commonly applied ones are presented in this paragraph. The simplest design structure is AB, which mirrors the natural therapy process with its evaluation and treatment periods. AB designs are indispensable when non-reversion behaviours and/or treatments with persistent effects are studied (e.g. learning processes). AB designs are also needed when treatment interruption is not advised due to clinical or social reasons and when time limitations restrict therapy continuation. According to Rabin (1981), AB designs are really useful in clinical settings as they mirror the natural therapy process which involves an initial assessment period (i.e., baseline) followed by an intervention period (i.e., treatment phase). However this design is not sufficient to demonstrate experimental control (Wampold & Furlong, 1981a). Multiple-baseline designs are the ones which replicate with delay and AB structure in different behaviours, settings or experimental units. This designs controls history effects as the intervention is introduced in different time moments. Other designs controlling for extraneous variables are ABA and ABAB, with the second one being preferred from an ethical point of view as it terminates with a treatment phase. In those designs an effective treatment is supposed to produce a change in the reversive behaviour only during the B-phases, while in the second A-phase a return to baseline levels is expected.

One of the features that distinguishes single-case data and makes their analysis controversial is serial dependence between the measurements of the same experimental unit. Recent surveys (e.g., Busk & Marascuilo, 1988; Matyas & Greenwood, 1991; 1997; Parker, 2006) have reported results suggesting that autocorrelation in usually present, in contrast with previous revision studies (Huitema, 1985; 1988). Several authors (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987) concur that even low and statistically non-significant levels of autocorrelation can critically increase the risk of Type I error when classical statistical tests are employed. Consequently, and taking into account the violation of the independence assumption, in the following sections the parametric tests commonly used for group studies (e.g., ANOVA) will not be discussed.

The main aim of the first part of this article is to present and review several techniques proposed for analysing N = 1 data, focusing on their application and on the results from previous investigations. In the second part of the study we centre on randomization tests and apply a new data simulation approach in order to study the statistical properties of this technique and obtain evidence on whether its application is to be advised or not.

Single-case data analysis

One of the most frequently applied methods for analysing N = 1 data is visual inspection. In an AB design, visual analysis requires a stable baseline (i.e., low behaviour variability in phase A) or a behaviour that shows a trend in a direction contrary to the one expected to be produced by the intervention (e.g., increasing number of cigarettes smoked during phase A when the intervention's objective is to decrease smoking). The treatment may produce different types of effects (see Figure 1): a) level change: abrupt increment or decrement in behaviour coinciding with intervention introduction; b) slope change: gradual increment or decrement; c) level and slope change. These effects can be produced with a delay (i.e., the behavioural change starts some time after the intervention has been initiated) or be decaying (i.e., return to baseline level during the treatment phase). Other characteristics that have to be taken into consideration are the variability within and across phases and the data overlap between phases (Ottenbacher, 1990). An effective treatment is supposed to produce rapid and maintained changes in behavioural rates, but delayed and extinguishing effects should not be overlooked. An advantage of visual inspection is that it does not require statistical expertise. Visual analysis has been proposed (Kratochwill & Levin, 1980) whenever large changes in level between phases are apparent and effect sizes somewhat greater than 2.0 appear to be sufficient (Matyas & Greenwood, 1990). With respect to that, it was claimed that this type of analysis, due to its relative insensitivity, ensures that only clinically relevant effects are detected (Parsonson & Baer, 1986). Empirical studies, however, have shown that treatment effect detection is affected by the presence of autocorrelation (Jones, Weinrott, & Vaught, 1978) and variability in data, often increasing the false alarm rates (Matyas & Greenwood, 1990). An additional drawback of visual inspection resides in the fact that no formal decision rules are available (Wampold & Furlong, 1981b).

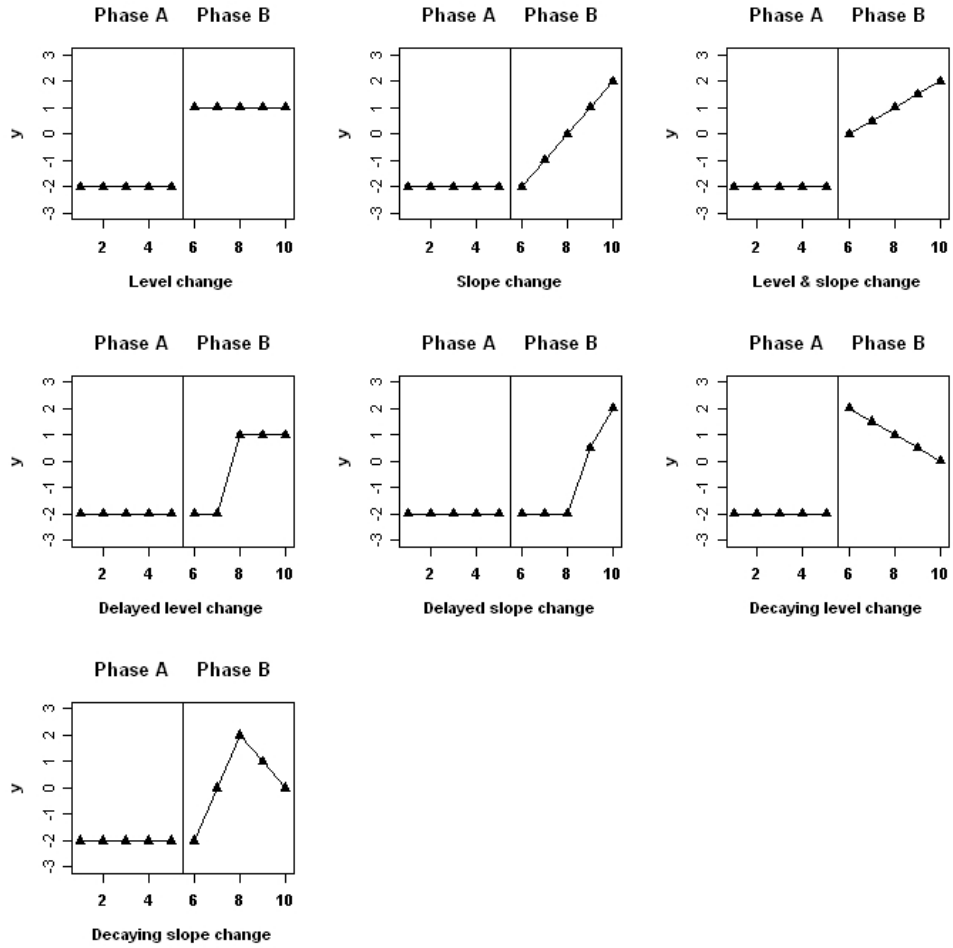

Figure 1. Idealised examples of different types of treatment effects in an AB design.

Another possibility for analysing N = 1 data is ARIMA (autoregressive integrated moving averages model), a procedure for interrupted time-series analysis proposed as a way of overcoming the autocorrelation problem (Crosbie, 1993; Kratochwill & Levin, 1980, Sharpley & Alavosius, 1988). An interrupted time series is a design in which one condition (e.g., baseline) is caused to cease by the introduction of another (e.g., treatment). The procedure described in Glass, Willson, and Gottman (1975, cited in Harrop & Velicer, 1985) involves the following steps: 1) Identify the model that fits empirical data – assess the pattern of autocorrelations and partial autocorrelations to determine the order of the

autoregressive (*p*), differencing (*d*), moving average (*q*) parameters. 2) Remove slope by differencing the data. 3) Determine the least squares estimates of the AR and MA parameters. 4) Remove autocorrelation using the parameters estimated in the previous step. 5) Apply the General Linear Model to determine if there is a level or slope change between the uncorrelated pre- and post-intervention scores. ARIMA is relatively more complex statistical technique and its application requires greater statistical expertise, while another limitation is the great number of observations required in order to accurately estimate the parameters (*p*, *d*, *q*) of the ARIMA model. Moreover, empirical results suggest that serial dependence may lead to inflation of Type I error rates, when positive, and to deflation, when negative (Greenwood & Matyas, 1990).

N = 1 data can also be analysed by means of randomization tests, although some concerns have been raised regarding this question (Cox & Hinkley, 1974). This procedure constitutes a specific way to determine the statistical significance of a treatment effect directly from data (Edgington, 1995), although no generalization to other experimental units is made due to the lack of random sampling (Edgington & Onghena, 2007). A randomization test is a permutation test which requires that some aspect of the design be randomized, but it does not involve any assumptions about population distributions, the nature of the data, or the kind of test statistic. In an applied setting a randomization test for an AB design could be used as described subsequently, although this is not the only possible random assignment procedure. 1) The researcher specifies his or her research hypothesis from which the null and the alternative hypotheses are derived. 2) The statistical significance level is chosen. 3) A test statistic sensitive to the effects expected is selected. 4) Design's length (*n*) is chosen and it is possible to set a minimum number of observations per each of the two phases. In the example in Figure 2 where $n = 30$ five measurement are preserved for phase A, and five more at the end of the series for phase B. 5) The starting point of the intervention is randomly chosen among all observation points and taking into account the restriction established in the previous step. For the data in Figure 2 there are 21 possible points for treatment introduction and the actually chosen intervention point is 13 leading to 12 measurements for phase A and 18 for phase B. 6) The test statistic is calculated for the actual data bipartition obtaining the outcome. 7) For each possible intervention point not chosen (i.e., 6, 7, …, 12, 14, …, 25, 26) the test statistic is calculated once again and, therefore, the intervention point rather than data is the aspect being varied. 8) The reference set, an equivalent to a sampling distribution, is obtained using all values of the test statistic. The division of the data made to obtain that

reference set should match the random assignment procedure actually used (Edgington, 1980b). 9) The value of the outcome is located in the reference set. 10) The p-value is calculated as the number of test statistics equal to or greater than the outcome, for a behaviour expected to increase. The randomization test is valid if the empirical probability of rejecting a true null hypothesis is no greater than alpha, that is, the nominal significance level (Edgington, 1980b; Hayes, 1996). On theoretical grounds it has been claimed that the presence of serial dependence is insignificant for randomization tests, emphasising the following reasons: a) the effects of autocorrelation are the same for all data permutations in presence of random assignment (Wampold & Worsham, 1986); b) the reference distribution is generated internally from the data themselves (Kratochwill & Levin, 1980); c) phase means can be used as they are approximately independent (Levin, Marascuilo, & Hubert, 1978); d) systematic trends in the data may affect the power of a statistical test, but have no effect on the ease of getting significant results when the null hypothesis is true (Edgington, 1980b). Empirical studies (e.g., Ferron, Foster-Johnson, & Kromrey, 2003; Ferron & Ware, 1995) concur with the latter statement showing that Type II errors are the main problem, while Type I errors are usually controlled.

In contrast with these statements and findings, the notion of randomization tests as a panacea has been questioned by the assertion that all hypothesis-testing methods rely on the independence and/or exchangeability of the observations (Good, 1994; Gorman & Allison, 1997). Empirical research has shown that significance probability values are underestimated for positive autocorrelated residuals, meaning that a researcher might be led to believe that a test is significant at the 0.05 level when in fact it is significant at a higher level (Gorman & Allison, 1997). There is also recent evidence for more-phased N = 1 designs that statistical significance of the *outcome* depends also on the specific data division (Manolov & Solanas, 2008; Sierra, Solanas, & Quera, 2005) and that Type I error rates are not always controlled. Consequently, the major part of the simulation study is focused on comparing the nominal and empirical Type I error rates for AB designs whose data presents different levels of autocorrelation (which violates the exchangeability assumption).

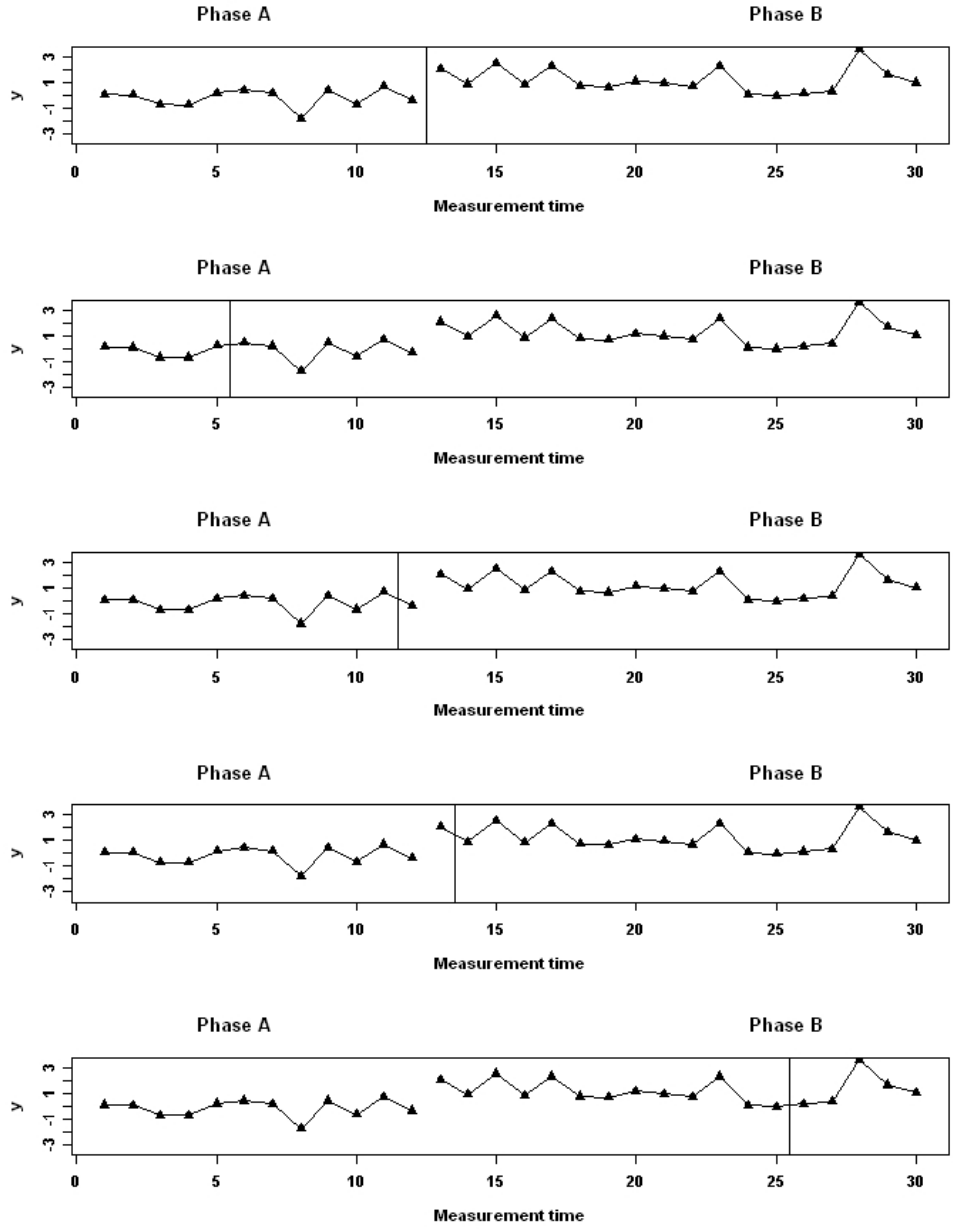

Figure 2. Randomization test applied to a n = 30 AB design. Actual intervention point: 13 (1st panel). Some other possible intervention points: 6 (2nd panel), 12 (3rd panel), 14 (4th panel), and 26 (5th panel). Data generation parameters: φ_1 = 0.3 and d = 0.8 (the latter applied only to points 13 to 30).

Randomization tests simulation study

In the present simulation study we applied the data-division-specific approach (Manolov & Solanas, 2008; Sierra, Solanas, & Quera, 2005) which implies that there can be a different reference set for each data division determined by the moment in which the intervention is introduced. The main objectives of the study were to compare the results with previous investigations in which no distinction is made between data divisions and so the same simulation parameters were employed. Our hypothesis was that the randomization test may prove to have inadmissible properties (high false alarm rates and/or high miss rates) for some intervention points in the case of AB designs.

METHOD

Following Onghena (1992), prior to data collection (in a simulation study it is rather "generation" than "collection") the following aspects have to be chosen:

a) The alternative hypothesis. Given that $H_0 : \mu_A \geq \mu_B$, $H_1 : \mu_A < \mu_B$.

b) The level of significance: alpha was set to 0.05.

c) The number of measurement times: AB designs with 30 and 40 observation points were studied. Following Edgington (1980a), in both cases a minimum of five measurements per phase is ensured in order to rule out the possibility of having too few (or no) treatment times for one of the conditions. Therefore, the number of intervention points admissible for the n = 30 design is 21 – the intervention can start at any point between 6 and 26, inclusive. The utilization of this design length is due to the fact that it was established (Edgington, 1980a) as the minimum necessary to obtain statistical significance beyond the 0.05 level. With the specified boundaries, for the n = 40 design there are 31 possible intervention points between 6 and 36, inclusive.

d) The random assignment procedure: in the simulation study the intervention point was systematically selected in order to obtain data-division-specific information. Therefore, we studied the effect of autocorrelation on Type I and Type II error rates for each intervention point in a systematic manner. It is a simulation procedure which has no relation to applied settings where the researcher should chose randomly the intervention point in order to validly use a randomization tests.

e) The test statistic: two of them were used. The first is expressed as $\bar{X}_B - \bar{X}_A$ and represents the difference between phase means (hereinafter, *MD*), previously used in various studies (Ferron & Onghena, 1996; Ferron & Ware, 1995). The second was Student's *t* (hereinafter, *ST*), calculated according to $\left(\bar{X}_B - \bar{X}_A\right) / \sqrt{s^2 / n_B + s^2 / n_A}$. Its inclusion is based on the importance that data variability seems to have when treatment effects should be detected. Previous studies (Sierra, Quera, & Solanas, 2000) have shown that taking variance into account is helpful when differences in mean level are evaluated.

Data generation

The research was conducted by developing FORTRAN 90 programs for generating data and performing further calculations. The first step was carried out according to exactly the same formula used in the studies with which a comparison is pretended (e.g., Ferron & Onghena, 1996; Ferron & Ware, 1995): $y_t = \varphi_1 * y_{t-1} + \varepsilon_t + d$. In this expression y_t represents the data point corresponding to measurement time t, y_{t-1} is the previous data point, φ_1 is the value of the lag-one autocorrelation coefficient, ε_t is the independent error following a normal distribution with mean zero and standard deviation equal to one, and d is the effect size.

The values of the error term were generated with the aid of NAG fl90 mathematical-statistical libraries (specifically, the external subroutines *nag_rand_seed_set* and *nag_rand_normal*).

The values chosen for the level of serial dependence (–0.6, –0.3, 0.0, 0.3 and 0.6) are commonly used in simulations (Ferron & Onghena, 1996; Ferron & Ware, 1995; Greenwood & Matyas, 1990) and cover the range of autocorrelation values presented in Parker (2006) – median negative autocorrelation of −0.2 and median positive autocorrelation of 0.42.

Following Ferron and Sentovich (2002), effect size was defined as the difference between phase means divided by the standard deviation of the error term in the baseline phase. Cohen (1992) has operationally defined small, medium and large effect size (when the difference between independent means is calculated) as 0.20, 0.50, and 0.80, respectively. The present study focuses on these values and complements them with others (1.10, 1.40, 1.70 and 2.00) used in previous studies (e.g., Ferron & Onghena, 1996; Ferron & Sentovich, 2002). In the Type I error rates study *d* was set to zero for both phases, while for the Type II error rates study the moment of introducing a non-zero value of *d* depended on the actual data bipartition (i.e., the actually chosen intervention point). For instance, the

Figure 2 data was generated by selecting 13 as intervention point and afterwards adding *d* to data points 13 to 30. Summing the effect size to all phase B measurements implies that an effective treatment is one that produces an immediate increment in the behaviour of interest.

The 20 numbers prior to the design's observation points were discarded in order to reduce artificial effects (i.e., to diminish the effect of anomalous initial values) (Greenwood & Matyas, 1990).

Simulation

The simulation for the study of Type I error rates consisted of 100,000 iterations for each combination of an intervention point and an autocorrelation coefficient value. For the power study, the same number of iterations was made for each combination of intervention point, autocorrelation coefficient and effect size values. The use of 100,000 iterations seems to ensure sufficient accuracy for the estimation of the statistical properties. This number of repetitions is greater than the one used in many previous studies (1,000 in Ferron & Ware, 1995; 5,000 in Ferron, Foster-Johnson, & Kromrey, 2003; 10,000 in Ferron & Onghena, 1996; 10,000 in Ferron & Sentovich, 2002; 40,000 in Sierra, Solanas, & Quera, 2005).

After data have been generated for different levels of φ_1 and *d* in accordance with the intervention point selected, the following steps took place: 1) calculation of the test statistic for the actual data bipartition, obtaining the *outcome*; 2) calculation of the test statistic for each data division; 3) construction of the reference set sorting the test statistics' values obtained for all possible intervention points; and 4) ranking the *outcome*, according to its position in the reference set.

Analysis

The basic data for the Type I error rates study were the proportions (out of 100,000 iterations) of each rank assigned to the outcome. Special attention was paid to extreme ranks and in order to obtain more stable estimates we averaged the relative frequencies of ranks 1 and 21, ranks 2 and 20, etc. The number of extreme ranks whose cumulative proportion reached values close to 0.05 (nominal alpha) without overcoming it were labelled as "critical" for null hypothesis rejection. In order to assess the importance of serial dependence in data, we compared the cumulative proportions of the critical ranks when $\varphi = 0.0$ (i.e., the cumulative proportion for independent data, hereinafter, CPID) with the cumulative

proportions of the same number of ranks when $\varphi \neq 0.0$. This comparison was carried out for each combination of data division and test statistic. The similarity between those proportions was evaluated by means of Bradley's (1978, cited in Robey & Barcikowski, 1992) stringent criterion: if the cumulative proportions for $\varphi = -0.6, -0.3, 0.3$, and 0.6 all fell within the interval CPID ± 0.1 * CPID, then the effect of autocorrelation was judged to be insignificant for the particular combination of intervention point and test statistic. Power analysis was performed only for those robust cases and it implied another step that consisted in computing the proportion of critical ranks assigned to the outcome for all combinations of intervention point, degree of serial dependence and effect size.

RESULTS

In this section, partial results will be presented, although more detailed information is available from the authors on request.

In the $n = 30$ design deviations from the robustness intervals were observed for intervention points 6, 7, 8, 23, 24, 25, and 26 both for MD (Figure 3) and ST (Figure 4). Those figures illustrate the underestimation and overestimation of Type I error rates occurring for the most extreme intervention points under autocorrelation. They also show how ranks' proportions vary across data divisions even in absence of serial dependence. In the $n = 40$ design with 31 possible intervention points and $\alpha = 0.05$, autocorrelation results in a not robust test for intervention points 6, 7, 8, 9, 33, 34, 35, and 36. For both design lengths this means that nominal and empirical Type I errors did not match for the aforementioned intervention points and there was an increased probability of false alarm rates or excessive conservativeness for positive and negative serial dependence, respectively.

For the n = 30 design, power for effect sizes of 0.2 and 0.5 (defined as "small" and "medium" by Cohen, 1992) is low, not greater than 0.08 and 0.14, respectively. Even when effect size is 2.0, the probability of rejecting a false null hypothesis is smaller than 0.63. For the extreme intervention points (6, 7, 8, 24, 25, and 26) the test has no power at the 0.05 level. Table 2 shows the power of the randomization test averaged across all data divisions including the one for it has zero power. Marascuilo and Busk (1988) suggest that if an $\alpha = 0.05$ decision rule cannot be generated, one can place the most extreme value of the test statistic (or the largest rank, in the case of the procedure studied here) in the critical region and reject the null hypothesis when that value (or rank) is obtained. Following this procedure,

power estimates greater than 0.80 can be obtained for intervention points 6 and 26, but the probability of Type I error is > 10%. For the n = 40 design the power estimates are lower than 0.1 and 0.2 for effect sizes of 0.2 and 0.5, respectively. When an effect size of 2.0 is present in data, only for a few combinations of intervention point, autocorrelation and test statistic does the power reach acceptable values of 0.8 or higher (Cohen, 1992). Positive autocorrelation of 0.6 has a discernible reducing effect on power across all intervention points. Greater power estimates were obtained for n = 40 than for n = 30, but 0.6 autocorrelation has the same effect of decreasing it. Ferron and Onghena (1996) comment that positive autocorrelation has a differential effect according to the type of design, specifically decreasing power where a random intervention point is used. According to these authors, in an AB design positive autocorrelation can mask the transition from one phase to another, as our results have also verified.

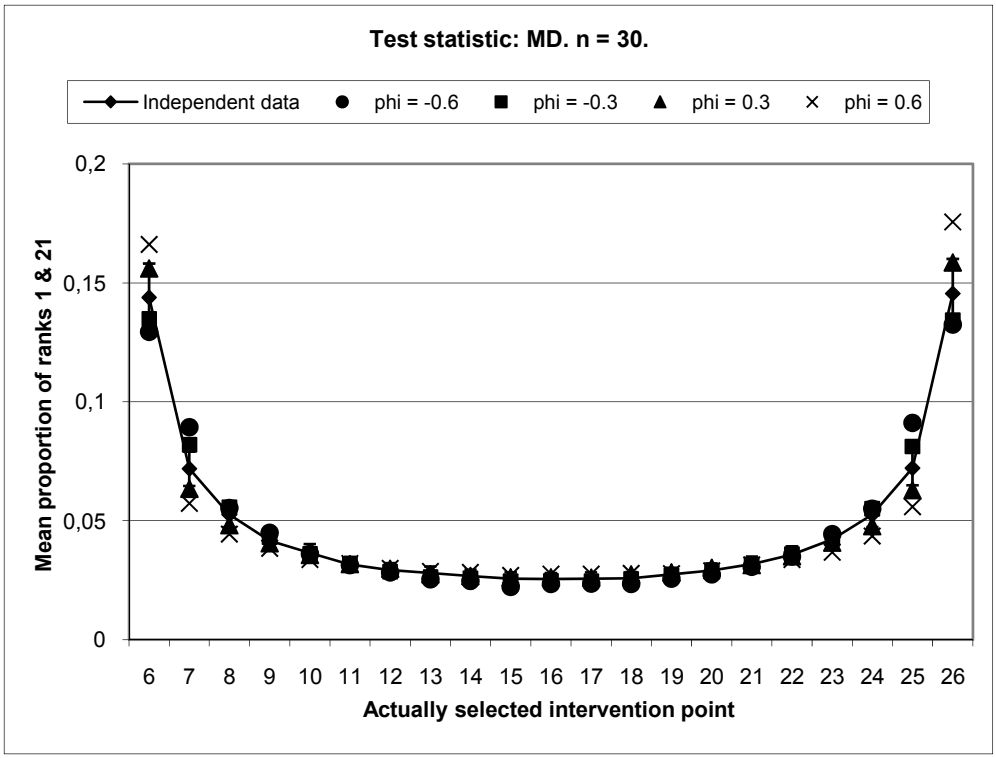

Figure 3. Mean proportion of ranks 1 and 21 assigned to the outcome computed through Mean Difference for each combination of admissible intervention point and level of autocorrelation. The deviations from the boundaries constructed about the φ_1=0.0 proportion indicate lack of robustness against the violation of the independence assumption.

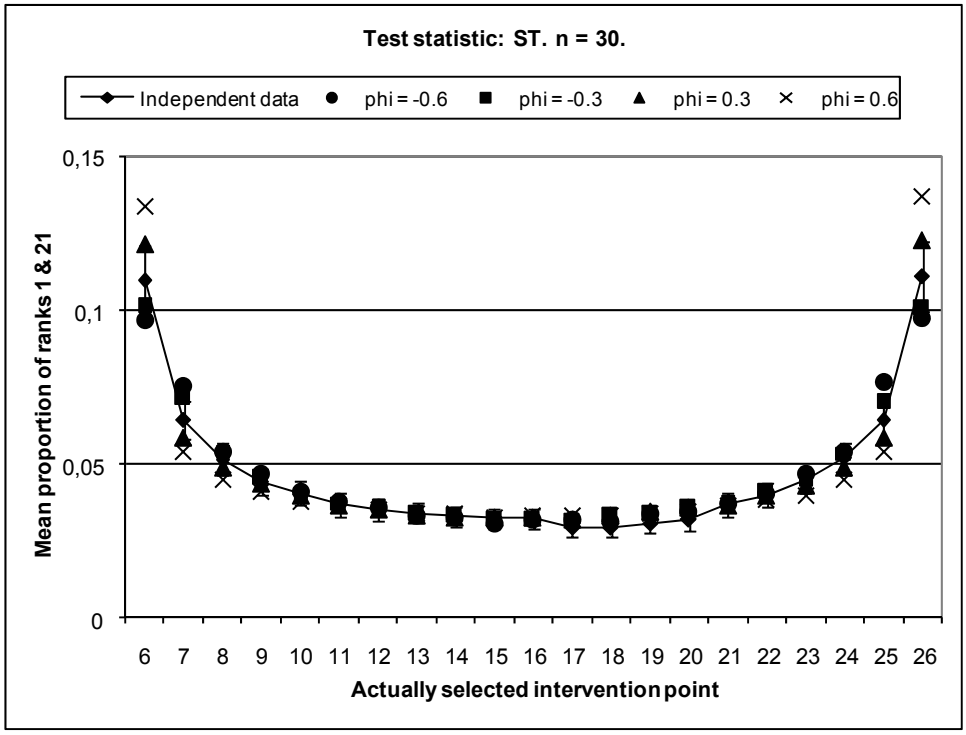

Figure 4. Mean proportion of ranks 1 and 21 assigned to the outcome computed through Student's *t* for each combination of admissible intervention point and level of autocorrelation. The deviations from the boundaries constructed about the $\varphi_1 = 0.0$ proportion indicate lack of robustness against the violation of the independence assumption.

DISCUSSION

Even for independent data series the probability of committing Type I errors clearly varies across the admissible data divisions. When $\varphi_1 = 0.0$, for both design lengths the Type I error rates are greater than 5% for the extreme intervention points, and the correspondence between nominal alpha and empirical probability of rejecting a true null hypothesis is important for preventing from statistical decision mistakes. Moreover, the effect of serial dependence ($\varphi_1 \neq 0.0$) is greater for the data divisions defined by those intervention points. A possible explanation resides in the fact the variances for phases with rather different sizes (i.e., short A phase and long B phase, or vice versa) are more unequal. These results do not concur with previous

findings (Ferron & Ware, 1995) and suggest that randomization test do not always control the Type I error rates and so their liberality cannot be ruled out.

Table 1. Mean power for all intervention points as a function of the autocorrelation level (φ_1) and the effect size (d) values; $\alpha = 0.05$ and $n = 30$. Power is equal to zero for the aforementioned nominal alpha for intervention points 6, 7, 8, 24, 25, and 26.

φ_1	Test statistic	$d = 0.0$	$d = 0.2$	$d = 0.5$	$d = 0.8$	$d = 1.1$	$d = 1.4$	$d = 1.7$	$d = 2.0$
-0.6	MD	0.0214	0.0330	0.0592	0.0954	0.1406	0.1922	0.2459	0.3193
	ST	*0.0261*	*0.0421*	*0.0783*	*0.1281*	*0.1880*	*0.2526*	*0.3148*	*0.3983*
-0.3	MD	0.0217	0.0350	0.0647	0.1069	0.1599	0.2200	0.2828	0.3456
	ST	*0.0260*	*0.0439*	*0.0845*	*0.1417*	*0.2111*	*0.2844*	*0.3562*	*0.4227*
0.0	MD	0.0221	0.0359	0.0668	0.1109	0.1675	0.2304	0.2972	0.3613
	ST	*0.0260*	*0.0445*	*0.0865*	*0.1466*	*0.2195*	*0.2972*	*0.3741*	*0.4432*
0.3	MD	0.0233	0.0356	0.0655	0.1072	0.1588	0.2153	0.2728	0.3275
	ST	*0.0258*	*0.0438*	*0.0852*	*0.1423*	*0.2100*	*0.2801*	*0.3480*	*0.4080*
0.6	MD	0.0217	0.0336	0.0575	0.0875	0.1193	0.1502	0.1762	0.1964
	ST	*0.0252*	*0.0418*	*0.0759*	*0.1179*	*0.1607*	*0.1972*	*0.2231*	*0.2381*

As far as sensitivity is concerned, for the $n = 30$ AB design a comparison with Ferron and Ware's (1995) results reveals a similar pattern of low power for $d = 1.4$. Our results indicate that even more evident effect sizes do not guarantee relatively small Type II error rates. Nonetheless, some additional comments on the effect sizes chosen for the power study are needed. Cohen (1992) sought to ensure that the 'medium' effect size represented an effect likely to be detectable by means of a careful visual analysis. However, Knapp (1983) found that visual judges show high agreement only when the intervention effect is greater than 2.0, a value that is quite different from Cohen's medium effect size of 0.5. Furthermore, a

survey by Matyas and Greenwood (1985, cited in Matyas & Greenwood, 1990), performed on articles extracted from the *Journal of Applied Behavior Analysis*, showed that the median effect size of AB panels with n ≥ 10 was 9.2, with percentile 25 equal to 4.9 and percentile 75 equal to 17.1. This comment is necessary to avoid conclusions being drawn on the power for medium effect sizes, as the concept 'medium' hardly represents any particular effect size. It might be also important to consider if Cohen's guidelines are appropriate for single-case designs.

The practical implication of these results is to show that, in presence of autocorrelation, false alarms and miss rates can be both too high when randomization tests are used to make decisions about the treatment applied to an experimental unit. The application of this statistical technique to AB designs is also limited by the high number of observations (30) required to obtain a p-value of 0.05 and by the fact that random selection of the moment in which to initiate intervention may not be feasible. While the first problem can be dealt with using more complex design structures (e.g., ABAB) that allow a greater number of admissible random assignments with shorter data series, the second one is inherent to randomization tests. Even when both these conditions are met the performance of the randomization test based on random selection of the intervention point is not satisfactory. Applied behavioural researchers should note that the results of our study recommend analysing data obtained from AB designs with the presented randomization test only when large effects are aimed to be detected, as smaller ones (i.e., effect sizes < 2.0) can be missed. Comparable performance is expected from visual inspection, but it constitutes a rather simpler technique, while software for performing randomization tests is still not widely available. Therefore, given the problems presented by techniques discussed in the first part of the article, the most parsimonious one ought to be recommended until a better solution is found for enhancing statistical decision making and facilitating applied researchers' labour.

The conclusions of the present study are restricted by the experimental conditions explored and its generalization to another set is not suggested. Only one type of design (AB) and only two design lengths (30 and 40) are studied. Moreover, the data simulated did not contain trends. However, these limitations are common to most studies focusing on similar tests (e.g., Ferron & Sentovich, 2002; Ferron & Ware, 1995).

Future randomization test simulation studies following the data-division-specific approach can be conducted for single-case designs following ABA, BAB, ABAB structures in which the points of change are randomly determined. Another possible line of research could focus on exploring rules that could guide visual analysts in their task.

RESUMEN

Problemas en las pruebas de aleatorización para diseños AB. Los diseños de caso único implican registrar la conducta de la misma unidad experimental. Las mediciones obtenidas suelen ser pocas debido a los costes temporales y también es probable que presenten dependencia serial. Estos problemas tienen que ser superados por las técnicas analíticas necesarias para mejorar la toma de decisiones estadísticas y clínicas. En la primera parte del artículo se discuten diferentes procedimientos para el análisis de diseños AB, presentando sus características y revisando resultados de estudios anteriores. Las pruebas de aleatorización son unos de los métodos estadísticos que se consideran apropiados debido a que parecen controlar las tasas de error Tipo I. En la parte experimental del artículo se utiliza una nueva manera de simular con el objetivo de analizar las propiedades estadísticas de las pruebas de aleatorización. Los resultados sugieren que la técnica no es siempre robusta contra la violación del supuesto de independencia y además presenta tasas de error Tipo II inaceptables. Teniendo en cuenta las evidencias disponibles, no parece existir una técnica óptima para el análisis de datos de N = 1.

REFERENCES

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.

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

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966-974.

Edgington, E. S. (1980a). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2*, 19-28.

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

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.

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

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

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

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

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167-178.

Good, P. (1994). *Permutation tests. A practical guide to resampling methods for testing hypotheses*. New York: Springer-Verlag.

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: Erlbaum.

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.

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.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods, 1*, 184-198.

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

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*, 253-294.

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.

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

Kratochwill, T. R., & Levin, J. R. (1980). On the applicability of various data analysis procedures to the simultaneous and alternating treatment designs in behavior therapy research. *Behavioral Assessment, 2*, 353-360.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). *N* = Nonparametric randomization tests. In T. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 167-196). New York: Academic Press.

Manolov, R., & Solanas, A. (2008). Randomization tests for ABAB designs: Comparing data-division-specific and common distributions. *Psicothema, 20,* 291-297.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10*, 1-28.

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 dependence 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.

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

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.

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.

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

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

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

Sierra, V., Quera, V., & Solanas, A. (2000). Autocorrelation effect on Type I error rate of Revusky's R_n test: A Monte Carlo study. *Psicológica, 21*, 91-114.

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.

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.

Wampold, B. E., & Furlong, M. J. (1981a). Randomization tests in single-subject designs: Illustrative examples. *Journal of Behavioral Assessment, 3*, 329-341.

Wampold, B. E., & Furlong, M. J. (1981b). The heuristics of visual inference. *Behavioral Assessment, 3*, 79-92.

Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8*, 135-143.

(Manuscript received: 6 September 2007; accepted: 28 February 2008)