Data-division-specific robustness and power of randomization tests for ABAB designs

ABSTRACT

This study deals with the statistical properties of a randomization test applied to an ABAB design in cases where the desirable random assignment of the points of change in phase is not possible. In order to obtain information about each possible data division we carried out a conditional Monte Carlo simulation with 100,000 samples for each systematically chosen triplet. Robustness and power are studied under several experimental conditions: different autocorrelation levels and different effect sizes, as well as different phase lengths determined by the points of change. Type I error rates were distorted by the presence of autocorrelation for the majority of data divisions. Satisfactory Type II error rates were obtained only for large treatment effects. The relationship between the lengths of the four phases appeared to be an important factor for the robustness and the power of the randomization test.

Key words: ABAB design, randomization tests, robustness, power

How to analyze single-subject data is a question yet to be answered. The different perspectives on the autocorrelation controversy have led to the proposal of a variety of techniques. It has been suggested that autocorrelation is a common feature in N = 1 data requiring the use of ARIMA to eliminate serial dependence before testing the treatment effect for significance (Jones, Weinrott, & Vaught, 1978). In contrast, Huitema and McKean (1998) have emphasized that the assumption of independence refers to residuals (or errors) and not to observations. If the autocorrelation between errors is nonsignificant, the authors recommend using widely known techniques like ANOVA (Huitema, 1985) and ordinary least squares when data series are short (Huitema, McKean, & McKnight, 1999). Despite that, the bias of autocorrelation estimators (Huitema & McKean, 1991) and the insufficient power of related significance tests in short series (Matyas & Greenwood, 1991) may lead to uncertainty regarding the size of the autocorrelation parameter and question the use of analyses based on the general linear model (Ferron, 2002).

Complementing these theoretical discussions, there is some empirical proof that in presence or even in absence of autocorrelation the performance of several analytical techniques can be deficient. For instance, the simplest and most frequently applied (as reported by Parker & Brossart, 2003) visual analysis does not seem to be as exempt from Type I errors (Matyas & Greenwood, 1990) as it was postulated (Parsonson & Baer, 1986). The time series analysis that was especially designed for dealing with autocorrelation

requires lots of observations, while biased autocorrelation estimators (Huitema & McKean, 1991) or lack of control of Type I error rates (Greenwood & Matyas, 1990) are other drawbacks it may present. Type I error rates have also been found to be problematic for ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983), the C statistic (Blumberg, 1984), the binomial test and the split-middle method (Crosbie, 1987).

As each of the previously mentioned techniques, randomization tests have their own advantages (Edgington & Onghena, 2007) and limitations (Kazdin, 1980). The appropriateness of their application to single-case data has on the one hand been seriously questioned (Cox & Hinkley, 1974), but on the other hand, randomization tests have been claimed to be useful for a great diversity of designs (e.g., Levin, Marascuilo, & Hubert, 1978; Levin & Wampold, 1999; Marascuilo & Busk, 1988; Onghena & Edgington, 1994; Wampold & Furlong, 1981). While some authors (e.g., Crosbie, 1987; Kratochwill & Levin, 1980, Wampold & Worsham, 1986) state on theoretical grounds that serial dependence is not a problem when randomization tests are used, empirical and nominal Type I error rates need to be compared in absence of treatment effect in order to accumulate evidence on test's performance (Hayes, 1996).

Most of the previous investigations focusing on the application of randomization tests to serially dependent single-case data (e.g., Ferron, Foster-Johnson, & Kromrey, 2003; Ferron & Onghena, 1996; Ferron & Ware, 1995; Lall & Levin, 2004) concur that Type II rather than Type I error rates are the main problem of the technique. For instance, for AB designs with 30

observations, power was less than .5 for an effect size of 1.4, while for designs with 32 observations following an ABAB structure power did not reach .6 for the same magnitude of effect (Ferron & Ware, 1995). Nevertheless, there is also evidence that Type I error rates are not always controlled when randomization tests are used (Gorman & Allison, 1996; Sierra, Solanas, & Quera, 2005).

In view of the surveys reporting that random assignment is not frequent in applied settings (Ferron & Jones, 2006), we considered it necessary to explore the performance of the randomization test under the influence of serial dependence for each possible data division, defined by the points of change in phase. This kind of information can be useful for applied researchers who are sometimes forced to choose the data division systematically – it shows them when to reject the null hypothesis maintaining the correspondence between nominal and empirical Type I error rates. The importance of data divisions can be expressed in terms of the length of each of the phases. Evidence on the variable performance of randomization tests when the only changing factor is the number of measurements in each phase has already been obtained for AB designs, showing that false alarm rates increase if either of the phases is a lot shorter than the other one (Manolov & Solanas, in press). Here, we want to study the importance of phase length for ABAB designs, extending the work of Ferron et al. (2003). Therefore, each data division studied preserves the phase order (i.e., ABAB) but varies the phase lengths (i.e., has different values of n_{A1}, n_{B1}, n_{A2}, and n_{B2}).

The randomization test studied here was proposed by Onghena (1992), and consists of randomly selecting the three points of change in phase (i.e., introduction, withdrawal, and re-introduction of the treatment). In the present paper, we are not advocating for a new randomization test, but we are rather studying and using a previous proposal in a different manner (i.e., systematic selection of the points of change in phase). Each particular data division is determined by a triplet of points of change and, consequently, each different data division will be called a "triplet" throughout this article. Each triplet is defined by the length of the four phases and will be identified by "$b_1.a_2.b_2$", where b_1 is the first data point for the first treatment phase, a_2 is the first data point for the second baseline phase, and b_2 is the first data point for the second treatment phase; the first data point for the first baseline phase (a_1) is obligatorily 1. Ferron et al. (2003) studied this type of randomization test applied to an ABAB design with n = 30, without distinguishing one specific data division (or triplet) from another. The authors found that when the procedure of selecting a triplet matches the method of permuting the data Type I error rates are controlled (.0468 for $\varphi = -.3$, .0484 for $\varphi = .0$, .0448 for $\varphi = .3$, and .0512 for $\varphi = .6$) and, hence, autocorrelation does not affect the ease of obtaining significant results when no effect is present in the data. However, when the triplet is systematically chosen, a practice we do not encourage as it contrasts with the inherent features of randomization tests, the statistical properties of the technique may vary across data divisions.

Therefore, the aim of the current study is to extend the contributions of published scientific literature and provide triplet-specific information by studying the statistical properties (i.e., robustness against the violation of the independence assumption and sensitivity) of the randomization test for the cases of independent and nonindependent data. The data-division-specific information implies that phase length is studied as a factor possibly related to the differential performance of the randomization test in terms of Type I and Type II errors.

Method

Selection of designs

An ABAB design with n = 30 is studied. The rationale behind this design length lies in the continuation of the work of Ferron et al. (2003) so that results can be compared. Following Edgington (1980), a minimum of five measurements per phase is established in order to rule out the possibility of having too few measurements for some of the phases. Applying the formula presented in Onghena (1992) with the current specifications, we obtain a totality of 286 possible triplets.

Data generation

Data were generated according to the following formula, employed previously in related investigations (e.g., Ferron & Onghena, 1996; Ferron & Ware, 1995; Matyas & Greenwood, 1990):

$y_t = \varphi_{1*} y_{t-1} + \varepsilon_t + d$, where:

y_t: data point corresponding to measurement time *t*;

y_{t-1}: data point corresponding to measurement time *t-1*;

φ_1: value of the lag-one autocorrelation coefficient;

ε_t: error term following N (0, 1);

d: effect size.

This expression was incorporated in FORTRAN 90 programs and the values of the error term were generated with the assistance of NAG *fl*90 mathematical-statistical libraries (specifically, the external subroutines *nag_rand_seed_set* and *nag_rand_normal*). The values chosen for the level of serial dependency (–.3, .0, .3, and .6) are commonly used in simulations (Ferron & Onghena, 1996; Ferron & Ware, 1995; Greenwood & Matyas, 1990) and are assumed to represent the range of autocorrelation present in behavioral data. Following Ferron and Sentovich (2002), effect size was defined as the difference between phase means divided by the standard deviation of the error term. For the study of the Type I error rates *d* was set to zero, while for the study of power the following values were used: .20, .50, .80, 1.10, 1.40, 1.70, and 2.00, as they are frequent in similar studies (e.g., Ferron & Onghena, 1996; Ferron & Sentovich, 2002). The studied effect was

an immediate and permanent change in level following the practice of previous research (e.g., Ferron & Onghena, 1996; Ferron & Sentovich, 2002).

Data generation involved simulating more numbers than the ones needed in order to discard 20 numbers between each pair of successive data series. This manipulation permitted reducing artificial effects (i.e., to diminish the effect of anomalous initial values) (Greenwood & Matyas, 1990) and ruling out the possibility of correlations between the last points of one series and the first points of the following one (Huitema, McKean, & McKnight, 1999).

Simulation

The simulation for the study of robustness consisted of 100,000 iterations (samples) for each combination of a triplet and an autocorrelation coefficient. The specific steps were: 1) successive selection of a triplet out of a list of all possible triplets commencing with data divisions with short first phase(s) (e.g., "6.11.16") and ending with data divisions with long first phase(s) (e.g., "16.21.26"); 2) systematic selection of the degree of serial dependence; 3) systematic selection of the effect size; 4) generation of the data; 5) calculation of each of the two test statistics for the actual data, obtaining the *outcomes*; 6) permutation of the data for all possible triplets and calculation of the test statistics for each data division; 7) construction of the randomization distribution sorting all values; and 8) ranking the *outcome*, according to its position in the randomization distribution.

Type II error rates were estimated for only a fraction of the 286 possible triplets – the ones for which the test did not seem to be affected by any of the degrees of serial dependence studied.

The use of 100,000 iterations seems to ensure sufficient accuracy for the estimation of the robustness and the power of a randomization test (Robey & Barcikowski, 1992).

Analysis

In an applied setting a particular educational or clinical study would start with a problem – excess or lack of a behavior – and a null hypothesis. Therefore, the researcher knows if the intervention is supposed to reduce an undesirable behavior or to enhance a positive behavior and would use a directional null hypothesis. The next steps to take if he/she is willing to apply a randomization test are described in detail by Onghena (1992). In the current study, the one-tailed null hypothesis was expressed as H_0: $\mu_A \geq \mu_B$ due to the fact that d was added to the measurements pertaining to phase B, simulating a treatment that increments a desirable behavior. The selected level of significance is 5% and two test statistics are used. The first one is expressed as $\bar{X}_B - \bar{X}_A$ and represents the difference between phase means (hereinafter, MD), previously used in various studies (Ferron & Ware, 1995; Ferron & Onghena, 1996). The second one is pooled variance *t* statistic (hereinafter, TS), calculated according to $\left(\bar{X}_B - \bar{X}_A\right) / \sqrt{s^2 / n_B + s^2 / n_A}$, which was included

as there is evidence that data variability is relevant when differences in mean level are to be evaluated (Sierra, Quera, & Solanas, 2000). As there are two test statistics computed on each data set, there are two different randomization distributions for each data division.

The first part of the assessment of the randomization test focuses on zero effect size (i.e., estimating Type I error rates). The *outcome* is compared to the randomization distribution and is ranked. The proportion of each of the 286 ranks is calculated. As alpha was set to .05 we sought to find, for each combination of data division and test statistic, the number of extreme ranks whose cumulative proportion is $\leq .05$. These ranks would represent the critical ranks for null hypothesis rejection. With a directional null hypothesis only one of the extremes of the randomization distribution would be used. For instance, when H_0: $\mu_A \geq \mu_B$, the null hypothesis would be rejected if the outcome is assigned some of the largest ranks, which correspond to greater positive difference between phase B and phase A. However, in order to reduce the effects of random fluctuations in the generated data (i.e., to base the estimation on a greater number of iterations), the proportions of critical ranks at both extremes were used. That is to say, we averaged the proportions of ranks 286 and 1 firstly, of ranks 285 and 2, secondly, of ranks 284 and 3 thirdly, and so on. After that, we summed those average proportions until the cumulative proportion became as close as possible to .05 without overcoming this value. This procedure showed that for the same triplet there can be a different

number of critical ranks according to the test statistic used, if we want the probability of committing a Type I error to be close to 5%.

In order to judge whether the effect of serial dependence in data is slight or important, we compared the cumulative proportions of the critical ranks when $\varphi = .0$ (i.e., the cumulative proportion for independent data, hereinafter, CPID) with the cumulative proportions of the same number of ranks when $\varphi \neq .0$. This comparison was carried out for each combination of data division and test statistic. With the objective to measure the similarity between those proportions, we used Bradley's (1978, cited in Robey & Barcikowski, 1992) stringent criterion. Out of the variety of existing criteria (e.g., the liberal and the intermediate), we chose the stringent one (which provides narrower intervals) as it rules out the possibility of too liberal or too conservative Type I error rates. This criterion can be viewed as a tool for marking the boundaries between what can be considered "similar enough" (i.e., robust) and what can be thought of as "too distant" (i.e., not robust). If the cumulative proportions for $\varphi = -.3, .3$, and .6 all fell within the interval CPID ± 10% * CPID, then the randomization test was qualified as robust for the particular combination of triplet and test statistic and the effect of autocorrelation was judged to be insignificant. Only those triplets for which the test was robust against the violation of the independence / exchangeability assumption were used in the power analysis (i.e., when $d \neq 0$).

As robustness was evaluated 286 * 2 times (for each combination of triplet and test statistic), it was possible to identify triplets for which the

randomization test is robust only for one of the test statistics. For instance, Table 1 shows how for triplet “8.13.26” with TS high positive autocorrelation makes the randomization test too conservative. On the other hand, Table 2 exemplifies a case in which autocorrelation has no critical influence on the Type I error rates, but in order to ensure an empirical rate of 5% the number of critical ranks is different. Among the “non-robust triplets” there were some for which the deviation from the robustness interval was small (Table 3), while for others the Type I error rate distortion due to autocorrelation was rather large (Table 4).

INSERT TABLES 1, 2, 3, & 4 ABOUT HERE

The power analysis was carried out in the following manner and only for the “robust triplets”: 1) identify the number of critical ranks for each combination of data division and test statistic; 2) use the randomization test to assign a rank to the *outcome* for each combination of degree of serial dependence and effect size; 3) count the number of times that the *outcome* has been assigned one of the corresponding critical ranks; 4) divide the value obtained in the previous step by 100,000 (the number of iterations) in order to estimate power.

Results

In this section, only part of the results will be presented in tabular format, although more detailed information is available from the authors upon request.

Table 5, containing Type I error rates averaged across all possible (286) triplets, shows that autocorrelation does not (in general) affect the probability of detecting a non-existent treatment effect. These results concur with previous findings on the correct performance of randomization tests applied in cases where random assignment is possible (Ferron et al., 2003).

INSERT TABLE 5 ABOUT HERE

Nonetheless, the data-division-specific Type I error estimates obtained showed that the randomization test is robust, under Bradley's stringent criterion, to the violation of the independence assumption for 52 triplets when using MD and for 51 triplets when using TS. The results do not suggest that one test statistic is better than the other in terms of controlling the Type I error rates. However, using one test statistic or another has two implications: a) the number of critical ranks needed to obtain a Type I error rate of .05 for $\varphi = .0$ may be different; and b) the randomization test for a specific triplet may be relatively unaffected by autocorrelation for one test statistic but not for the other. Therefore, the performance of the randomization test is not independent from the test statistic used.

Each of the triplets for which the randomization test was found to be insensitive to serial dependence has its number of critical ranks that guarantee a Type I error rate approximately equal to .05. For instance, if an applied researcher using MD as a test statistic has chosen triplet "6.13.21", he/she

should reject the null hypothesis if the outcome is assigned one of the 15 most extreme ranks. In case TS is the test statistic, the 14 most extreme ranks ought to be used for null-hypothesis rejection for the same triplet. Then the probability of committing a Type I error would be approximately .05, if $-.3 \leq \varphi \leq .6$. Using another decision rule does not guarantee the matching between nominal and empirical false alarm rates.

With Table 6 we explore phase length, defined by the n_{A1}, n_{B1}, n_{A2}, and n_{B2} values, in relation to Type I error rates for all 286 triplets. An inspection of those rates allows identifying the phase length pattern of the "robust" triplets, on one hand, and the pattern of the data divisions most affected by serial dependence, on the other. We arbitrarily labeled phases with length 5 to 8 as "short" (S) and phases with more than 8 observations as "long" (L). It appears that positive autocorrelation leads to a more conservative test when the triplet has approximately equally long phases (i.e., the SSSS pattern), when it has short baseline and long treatment phases (i.e., the SLSL pattern) and vice versa (i.e., the LSLS pattern). When the first and the last phases are shorter (SLLS pattern), positive autocorrelation is associated with a more liberal test. The performance of the randomization test was found to be more satisfactory for triplets with one long and three short phases and for data divisions with short second and third phases (LSSL).

INSERT TABLE 6 ABOUT HERE

An additional analysis of the phase length pattern of the "robust" triplets can be found in Table 7. The most distinguished pattern is LSSL, which seems the case for which serial dependence produces less distortion. Concurring with the previous table, the most affected data divisions were the ones with four equally long phases (SSSS) and the ones with short phases in either of the conditions (SLSL and LSLS).

INSERT TABLE 7 ABOUT HERE

The power study was performed for the combinations of triplet and test statistic for which the test was judged to be robust, as the lack of control of Type I error rates presented by the remaining combinations renders meaningless the estimation of Type II error rates for them. The mean power across the 52 MD-triplets and the 51 TS-triplets can be found in Table 8. In terms of sensitivity, as was the case for robustness, none of the test statistics outperformed the other. It is evident that Type II errors can be excessively probable unless the treatment effect is rather large (i.e., d = 1.7). Smaller effects are likely to be missed by the technique, as power estimates for the effect sizes labeled by Cohen (1992) as "small" (d = .2), "medium" (d = .5), and "large" (d = .8) are lower than .40.

INSERT TABLE 8 ABOUT HERE

However, power varies across triplets. Table 9 presents the power estimates for the "robust" triplets when d = 2.0, grouped according to the phase length pattern. Higher power was found for triplets with one long phase (SSLS, SLSS, and LSSS) and with two initial long phases (LLSS). Complementarily, lower power estimates were associated with data divisions in which A_2 and/or B_2 are the phases containing more data points (SSSL, SSLL). An evident general tendency is that high positive autocorrelation is related to lower sensitivity.

INSERT TABLE 9 ABOUT HERE

Discussion

The triplet-averaged results presented in Table 5 show that for the correct application of the randomization test (i.e., selecting randomly the data division) the empirical Type I error rate is approximately equal to alpha, concurring with the result reported by Ferron et al. (2003). However, whenever the randomization test is applied systematically, it is necessary to distinguish one triplet from another and to obtain information about the influence of serial dependence for each data division. The present simulation research shows that the estimated empirical Type I error rates for a nominal value equal to .05 clearly vary across the different data divisions. Phase length appears to be an important factor in terms of robustness and power. The

randomization test in its systematic application performs better for data divisions in which the only long phase is one of the first three and for data divisions with two long phases – the first and the last. From a clinical perspective an LSSL pattern may be suitable, as it would allow sufficient time for assessing the existing situation and establishing a stable baseline. It would also have the advantage of a long treatment phase at the end of the professional-client relationship.

The most important consequence for single-case analysts is that, once a specific triplet has been systematically selected, the randomization distribution can be different from that of other triplets. Therefore, for the same risk that the researcher is willing to assume (i.e., for the same nominal alpha) the number of critical ranks is different. That is to say, different rank values may be associated with the same nominal Type I error rate. Rejecting the null hypothesis when the outcome is assigned one of the 14 most extreme ranks (as the 286 * .05 = 14.3 calculation suggests) would lead to Type I errors in 5% of the cases only when all ranks are equally probable. However, for some of the triplets it is easier to obtain extreme ranks and a rejection rule based on 14 ranks would make Type I errors more frequent. For other triplets it is more difficult to obtain extreme ranks and, hence, the probability of committing Type I errors would be smaller than .05 but this would also lead to a decrease in power. Therefore, in order to control Type I error rates and not to lose power, the ranks which lead to null hypothesis rejection ought to be the adequate ones for the specific data division systematically chosen. If applied

researchers cannot randomly choose the triplet, their systematic selection can be limited to the triplets for which the randomization test was found to be less affected by autocorrelation. The results presented here give more information to applied researchers and permit improving the planning of single-case studies. They can choose a "robust" triplet that matches best their specific case out of the list presented in Appendix 1. The rule for rejecting the null hypothesis of no treatment effect would be determined by the number *k* of critical ranks associated with the triplet chosen. If the test statistic is assigned one of the *k* most extreme ranks, then the applied researcher would have evidence of the effectiveness of the intervention. The R codes presented in Appendices 2 and 3 (for MD and TS, respectively) perform all necessary calculations leading to the statistical decision and only require that the applied researcher enters the data obtained and the three points of change in phase actually used.

As regards Type II error rates, the power estimates obtained in the present study are similar to the ones obtained by Ferron and Ware (1995) and it was possible to identify the same tendency of less sensitivity for greater degrees of autocorrelation. Comparing our ABAB-results with the ones obtained for other types of designs containing 30 observation points, we found similar power estimates as in multiple-baseline designs (Ferron & Sentovich, 2002) and lower sensitivity than in six-phase designs (Ferron & Onghena, 1996).

In relation to the test statistic that can be used in the randomization test, it has to be adverted that the distorting effect of autocorrelation is not

independent from the test statistic used. Out of the two test statistics studied it cannot be claimed that one is more recommended than the other. The evidence obtained on MD and TS shows that when choosing the triplet and the decision making rule, the test statistic employed has to be taken into consideration.

Finally, we have to address the question of the adequacy of the triplet procedure studied as an analytical technique for single-case designs. It should be highlighted that randomization tests are designed to be applied in conditions of random assignment and the validity of their "systematic" use is questionable. Moreover, the usefulness of the randomization test studied is limited by the number of measurements required, which does not seem to correspond to the average series' length reported by Huitema (1985). On the other hand, for the median negative and positive autocorrelations found by Parker (2006), −.20 and .42 respectively, the procedure studied controls false alarm rates for some of the triplets and detects only powerful treatments, whose effects go beyond statistical significance and have potentially greater probability to be clinically meaningful.

The results of the present investigation should be considered with prudence, as it has centered only on one type of treatment effect (immediate and permanent change in level) and on one specific series length. Additionally, the information provided here may not be exactly accurate for randomization tests using test statistics other than MD and TS. The Type I and Type II error estimates presented here are specifically relevant for the occasions in which applied researchers are forced to select the triplet systematically and further

generalization is not advised. Finally, a last limitation stems from the fact that a random sampling model was used to generate the simulated series. Under the null hypothesis of a pure random assignment model, there are no other scores than the observed scores, and it can be shown that the randomization test is, by definition, perfectly valid in that case (Edgington & Onghena, 2007).

Future research may focus on assessing the statistical properties of randomization tests applied to designs with fewer phases (i.e., ABA and BAB) as they permit using less observation points and are, therefore, more feasible. The studies on those types of designs can be based on the methods previously applied and also on the one followed in the current study, in order to obtain evidence on the performance of the analytical techniques in settings where random assignment is possible, and also in cases where it is not. Additionally, more attention should be paid to phase lengths in order to explain why certain patterns enhance the performance of randomization tests, while others distort it.

References

Blumberg, C. J. (1984). Comments on "A simplified time-series analysis for evaluating treatment interventions". *Journal of Applied Behavior Analysis, 17*, 539-542.

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

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

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

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers, 34*, 324-331.

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., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66-81.

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.

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.

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.

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

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least squares intervention models. *Psychological Methods, 3,* 104-116.

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 dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kazdin, A. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics, 5,* 253-260.

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.

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

Levin, J. R., 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.

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly, 14*, 59-93.

Manolov, R., & Solanas, A. (in press). Problems of the randomization test for AB designs. *Psicológica.*

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.

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

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

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.

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.

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.

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.

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.

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.

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

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