Complete title:

Random assignment of intervention points in two phase single-case designs: Data-division-specific distributions

Authors and affiliations:

Antonio Solanas[1], Vicenta Sierra[2], Vicenç Quera[1], and Rumen Manolov[1]

[1] University of Barcelona

[2] ESADE-Ramon Llull University

Running head:

Data-division-specific randomization distributions

Contact information:

Address correspondence to Antonio Solanas, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron 171, 08035-Barcelona, Spain or mail: antonio.solanas@ub.edu.

Footnote:

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* grants SEJ2005-07310-C02-01/PSIC and SEJ2005-07310-C02-02/PSIC, and the *Generalitat de Catalunya* grant 2005SGR-00098.

SUMMARY. The present study explores the statistical properties of a randomization test based on the random assignment of the intervention point in a two-phase (AB) single-case design. The focus is on randomization distributions constructed with the values of the test statistic for all possible random assignments and used to obtain *p*-values. The shape of those distributions is investigated for each specific data division defined by the moment in which the intervention is introduced. Another aim of the study consisted in testing the detection of inexistent effects (i.e., production of false alarms) in autocorrelated data series, in which the assumption of exchangeability between observations may be untenable. In this way, it was possible to compare nominal and empirical Type I error rates in order to obtain evidence on the statistical validity of the randomization test for each individual data division. The results suggest that when either of the two phases has considerably less measurement times, Type I errors may be too probable and, hence, the decision making process to be carried out by applied researchers may be jeopardized.

Key words: Randomization tests, AB single-case design, random intervention point

Single-case designs are useful in psychological and educational research, as they permit examining the effects of a treatment over time for an individual subject or a group taken as a whole. An important distinction to be made is between single-case designs and case studies in terms of experimental rigor (Backman & Harris, 1999). Regarding data analysis of single-case designs, agreement among researchers has been found to be low (Ferron & Ware, 1995). The main concern commonly arises from the autocorrelated errors that are often assumed to exist in behavioral data. Autocorrelation (also referred to as "serial dependence") concerns the existence of a relationship (i.e., lack of independence) between measurements sequentially ordered in time. When an applied study involves the registration of a single experimental unit, it is likely that its behavior at one moment is related to its previous behavior. Although it has been advocated that conventional statistical methods can be properly employed for analyzing single-case designs data (Huitema, 1985), empirical evidence suggests that the presence of serial dependence can be problematic for several analytical techniques. As regards visual inspection of graphed data, as the most commonly applied method for single-case data analysis (Parker, Cryer, & Byrns, 2006), serial dependence disturbs agreement between statistical and visual inference (Jones, Weinrott, & Vaught, 1978) and increases Type I error rates (Matyas & Greenwood, 1990). In relation to parametric statistical tests, *t*-test for level does not perform properly in presence of serial dependence (Greenwood & Matyas, 1990), as Type I empirical error rates are distorted, similar results being obtained for ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983). Another strategy for analyzing behavioral data consists in statistically modeling the dependencies in the error structure, but this requires phase lengths that are uncommon in single-case designs (Ferron & Ware, 1995; Greenwood & Matyas, 1990).

Permutation or randomization tests have also been proposed as a way of statistically analyzing single-case experiments (Edgington, 1967; Edgington & Onghena, 2007). These permutation methods require some characteristic of the design to be randomized and a test statistic sensitive to the expected effect of the intervention to be chosen. Random assignment is an essential condition for a randomization test to meet internal and statistical validity (Edgington, 1980a). After conducting the experiment, the researcher computes the test statistic and determines statistical significance by locating where the obtained test statistic falls within the permutation or randomization distribution. This randomization test allows researchers to test both change in level and change in slope, the permutation procedure being identical apart from the definition of the statistic of interest (Wampold & Furlong, 1981).

Randomization tests are supposed not to make any assumption about the shape of distributions and, as a consequence, have been considered distribution-free (Edgington, 1980a; Marascuilo & Busk, 1988). However, comparing average performance in different experimental conditions can be obstructed by differences in variance (Gorman & Allison, 1997). Moreover, the precision of the results obtained by randomization tests depends on the exchangeability of observations (Good, 1994). That is, data permutations are only suitable when measurements' order does not influence the value of the test statistic (Good, 1994; Randles & Wolfe, 1979). In cases where one observation is related to the previous one (i.e., when series are autocorrelated), the exchangeability of data points is dubious, as the sequence in which they are obtained is relevant (Good, 1994).

The exchangeability of observations is important for preventing Type I error rates distortions and so for ensuring the validity of the randomization test. A statistical test is said to be statistically valid when the probability of committing a Type I error is less

than or equal to nominal alpha set by the applied researcher prior to conducting the experiment (Edgington, 1980a; Hayes, 1996). The need for the exchangeability assumption has been recognized in randomization tests, although it has often been established as the requirement for independence among data or nonautocorrelated errors (Levin, Marascuilo, & Hubert, 1978; Marascuilo & Busk, 1988). Regarding the serial dependence and statistical validity of randomization tests, it has been stated that these tests overcome autocorrelation problems (Crosbie, 1987; Levin et al., 1978; Wampold & Worsham, 1986). Nevertheless, some preliminary results of simulation studies have shown that randomization tests do not control Type I error rates if data are autocorrelated (Gorman & Allison, 1997). Recently, other simulation studies have found that at least some randomization tests do not control Type I error rates in the presence of serial dependence (Ferron, Foster-Johnson, & Kromrey, 2003; Sierra, Quera, & Solanas, 2000; Sierra, Solanas, & Quera, 2005).

The AB single-case design is the most basic form of single-case phase design (see Bulté & Onghena, 2008, for a discussion on phase and alternation designs). It involves a succession of two experimental conditions – a baseline or control phase (designated by A) is followed by a treatment phase (B) which lasts until the end of the study without being withdrawn. An effective treatment implies that the level of behavior during phase B deviates from the projected level of baseline performance (Kazdin, 1978). The fact that there is only one change in the experimental conditions implies that internal validity is not guaranteed. History, maturation, testing, and instrumentation effects are common examples of threats to internal validity. Nevertheless, the AB single-case design is often used in applied research, both in clinical and nonclinical settings, especially for nonreversive behaviors, in spite of its drawbacks. That is why the present study focuses on a randomization test for analyzing the data resulting from the AB single-case design.

Random assignment of an intervention point

Let us take for example a 30-point AB single-case design, in which the time of introduction of the intervention is randomly determined prior to collection of the data (Edgington, 1975). The selection of the intervention point determines the lengths of both phases, assigning the measurement times previous to that point to phase A and the remaining ones to phase B. The random choice of the point of intervention must be restricted to guarantee that neither of the two phases, A and B, has an excessively small number of data points – for instance, Edgington (1980b) suggests a minimum of five measurement times per phase, that is, $k = 5$. Therefore, considering the series' length ($n = 30$), the intervention point could be randomly selected from the set of integers ranging from $p = 6$ to $p = 26$, p_0 being used to denote the randomly chosen intervention point. Thus, there are 21 possible assignments (denoted by q) of the intervention point. q can be obtained through the following expression $n-2k+1$, which in the example presented is equal to $30-2(5)+1 = 21$. The experimenter could randomly select one of the following bipartitions, where the first and second numbers in each parenthesis respectively correspond to the number of measurements in phases A and B: (5, 25), (6, 24), …, (24, 6), (25, 5). It should be noted that (5, 25) is equivalent to $p_0 = 6$, (6, 24) to $p_0 = 7$, and so on. Note that any bipartition is equally probable before randomly choosing the intervention point. After randomly selecting the intervention point p_0, the experiment is carried out. The value of the statistic that is relevant and sensitive to the purpose of the research is firstly calculated for the observed data, that is, taking into consideration the actually selected intervention point and the *outcome* (denoted by d_0) is obtained. The same test statistic is then computed for all possible random assignments

of the point of intervention, which are represented by the remaining 20 (not selected) data bipartitions. The randomization distribution is then constructed by sorting all 21 possible values of the statistic (denoted as d_6, d_7, …, d_{26}) in an ascending order. Then, by means of the order statistic, the values of the statistic can be ordered. Thus, $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(21)}$. The value of the statistic for the data at hand (d_0) is located in the randomization distribution. It has been assumed that the statistical significance associated with the outcome is the proportion of test statistics as large as or larger than the obtained value (Edgington, 1980b; Wampold & Furlong, 1981). At least 20 possible intervention points would be required to allow for the possibility of statistical significance at the .05 level. When $q = 21$, the minimal possible p-value is 1/21 = .0476.

This way of determining the statistical significance of the outcome is founded on the common randomization distribution – a procedure that mixes all possible intervention points to generate the randomization distribution independently of the specific random intervention point that was selected by chance. The abovementioned procedure of obtaining p-values is based on the idea that the randomization distribution follows a discrete uniform or rectangular distribution for all admissible randomly chosen intervention points. Evidence suggests that mixing all possible data division actually leads to a uniform randomization distribution (Manolov & Solanas, 2008). However, when randomization distributions are investigated for each data division, shapes different from the rectangular appear (Manolov & Solanas, 2008; Sierra et al., 2005). Shapes' variation is reflected in disparity in Type I error rates. Therefore, the statistical significance of the outcome ought to be determined individually for each specific data division (i.e., using data-division-specific randomization distributions).

The idea subjacent to the common randomization distribution can be expressed by Equation 1:

$$\Pr\left(d \le d_0\right) = \frac{card\left\{d_{(i)} \le d_0\right\}}{n - 2k + 1}, \quad (1)$$

where *Pr ()* corresponds to the *p*-value associated with the outcome, *d* denotes the test statistic of interest (e.g., mean difference between phases A and B) and *card {·}* denotes the number of set elements.

On the other hand, the idea underlying data-division-specific randomization distributions can be expressed by Equation 2:

$$\Pr\left(d \le d_0 \mid p_0\right) = \frac{card\left\{d_{(i)} \le d_0 \mid p_0\right\}}{n - 2k + 1}, \quad (2)$$

where the only difference with respect to Equation (1) is that the *p*-value (*Pr*) and the number of set elements (*card {·}*) are conditional to the intervention point, as the term "$\mid p_0$" denotes.

After randomizing the intervention point, the way in which the specific design will be carried out is absolutely determined. That is why the proper randomization distribution is that associated with the specific intervention point that was randomly chosen. Then, the data-division-specific randomization distribution is the appropriate distribution to determine the statistical significance, and not the common randomization distribution (Sierra et al., 2005).

The main aim of the present study was to explore if the variation of distribution shapes and Type I error rates, in independent data series, across data divisions found for ABAB designs (Manolov & Solanas, 2008; Sierra et al., 2005) is also applicable to two-phase designs. The influence of autocorrelation for each specific intervention point was also to be tested, while additional objectives consisted in proposing an explanation of the results and showing their practical importance for applied researchers.

Method

A Monte Carlo simulation was conducted to estimate data-division-specific randomization distributions and to determine the effect of autocorrelation levels on the statistical decision-making process when the method of randomization involves the random assignment of an intervention point within the series of measurement times. The AB single-case design consisted of 30 observations, and at least five observations in each phase were planned, leading to 21 possible data bipartitions (Wampold & Furlong, 1981).

Data Generation

FORTRAN programming was used to generate AB single-case designs with 30 measurement times each and autocorrelations (φ_1) of −.9, −.6, −.3, .0, .3, .6, and .9. These values are common in randomization tests simulations (e.g., Ferron et al., 2003; Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995). The program then computed values of the statistic of interest and its randomization distribution. In the data-generation process, NAG mathematical-statistical libraries were used to generate normal random values for the error term of the autoregressive model and to set the initial seeds for data simulation, respectively. Data were generated according to Equation 3:

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t \quad (3)$$

where y_t and y_{t-1} are data points corresponding to measurement times t and $t-1$, φ_1 is the first order autoregressive parameter, and ε_t are $N(0,1)$ random variables. For each call to the NAG libraries, 130 data (ε_t) were generated, and the first 100 were discarded to reduce artificial effects (Greenwood & Matyas, 1990), that is, to attenuate as far as possible the effect of anomalous initial values or seeds of the pseudorandom generator and to stabilize the series. The remaining 30 data points were used in the analysis.

According to Robey and Barcikowski (1992), the number of iterations in a simulation needed for detecting deviations from the exact Type I error rates under the strong criterion $\alpha \pm 1/10\ \alpha$, a Type I error rate $\omega = .01$, and a prior power $1 - \beta = .9$, is 29,600. The forty thousand iterations used in the present study amply satisfy those criteria.

Test statistics

Two statistics were computed for each simulated data series. One of them was the difference between the mean for phase A and the mean for phase B, called thereafter *Statistic 1*. *Statistic 2* was computed as presented in Equation 4:

$$t = \frac{\left(\overline{x}_A - \overline{x}_B\right)}{\sqrt{s^2 / n_A + s^2 / n_B}} \qquad (4)$$

where s^2, n_A and n_B, respectively, correspond to the pooled estimation of the variance, the number of observations in phase A and the measurement times in phase B. Both statistics were calculated, since empirical Type I error rates could depend upon how the statistic was defined. While *Statistic 2* takes into account phase lengths and variability,

Statistic 1 does not. Data-division-specific randomization distributions for *Statistic 2* might be more similar to the discrete uniform distribution than those for *Statistic 1*.

Simulation

The steps in the simulation were as follows: data points were generated according to Equation (3) for a given φ_1 and a random intervention point; the outcome was computed for the data series using both *Statistic 1* and *Statistic 2*; admissible intervention points were permuted and the statistic is computed for each; values of the statistic are sorted to obtain the exact randomization distribution; the outcome was located in the randomization distribution and its rank (i.e., an integer between 1 and 21) is obtained. The abovementioned steps were repeated 40,000 times. These steps were repeated for each autoregressive parameter value and each possible random intervention point. In total, 147 experimental conditions were investigated, the combination of 21 possible random intervention points and 7 autocorrelation values.

Results

Randomization distributions in absence of autocorrelation

Table 1 shows summary statistics for data-division-specific randomization distributions as a function of the randomly selected intervention point and the test statistics.

TABLE 1 ABOUT HERE

Since data series were generated for $\varphi_1 = .0$, the exchangeability condition was met. While the mean of the ranks associated with the outcome was close to 11 for all data-division-specific randomization distributions, the variance of those ranks ranged from 27.099 to 55.574 according to the intervention point. The mean of ranks corresponded to the mathematical expectancy expected in case the data-division-specific randomization distribution follows a discrete uniform distribution. However, the variance expected for that distribution shape, $(21^2 - 1)/12 \approx 36.667$, did not approximate the dispersion values obtained for all data bipartitions. As regards the two test statistics used, the main difference between them is that for some intervention points *Statistic 1* presented greater variability, while for others it was *Statistic 2*. All data-division-specific randomization distributions showed an evident symmetry for both statistics, and that is why the mean ranks were close to the mathematical expectancy for each random intervention point. The kurtosis for a discrete uniform distribution ranging from 1 to 21 is approximately equal to -1.202, but the simulation study showed that, in general, data-division-specific randomization distributions had different kurtosis values. The two statistics also showed differences in their kurtosis values. Furthermore, considering the empirical Type I error rates, those values corresponding to the ranks 1 and 21 did not match 1/21 = .0476 (see Figure 1), which is the expected value for a discrete uniform distribution. Therefore, data-division-specific randomization distributions are not uniformly distributed for independent data series.

FIGURE 1 ABOUT HERE

In contrast, if no distinction is made regarding the intervention points and if the common randomization distribution is considered, all summary statistics resemble what

is expected for a discrete uniform distribution (see Table 1). Then, since data-division-specific randomization distributions should be used to determine statistical significance, it can be concluded that, in general, the Type I error rate associated with the most extreme values of the statistic of interest was not equal to .0476. The minimal value of the Type I error rate depended upon the intervention point, respectively ranging from .1442 to .0256 and from .1100 to .0324 for *Statistic 1* and *Statistic 2*. It should be noted that, regardless of how the statistic was computed, empirical Type I error rates were less than .05 for p_0 ranging from 9 to 23. Then the statistical test was valid, when the null hypothesis was true, at the level of statistical significance equal to .05 for any value from the set of integers ranging from 9 to 23. That is to say, when an applied researcher uses the randomization test to obtain evidence of treatment effectiveness, there is an increased risk of false alarms (i.e., detecting inexistent effects) if the intervention is introduced at measurement times 6, 7, 8, 24, 25, and 26.

Figure 2 shows the estimated mass probability for each possible rank associated with the outcome for *Statistic 1*. It is apparent that the distribution of the ranks depended upon the random intervention point. The mass probability function was approximately *U*-shaped for $p_0 = 6$, shows two modes at the ranks 6 and 16 for $p_0 = 11$ and had one mode at the center for $p_0 = 16$. It should be noted that the data-division-specific randomization distributions were symmetric and their variance values were reduced as the random intervention point approached 16. A comparison between Figures 2 and 3, representing results for *Statistics 1* and *2* respectively, reveals similar distribution shapes between both test statistics.

FIGURES 2 AND 3 ABOUT HERE

The effect of autocorrelation

The results described above suggest that empirical Type I error rates are equal or inferior to nominal ones (i.e., statistical validity is ensured) for the majority of data divisions – when the intervention point is between 9 and 23, both inclusive. For those cases, it was important to know whether the presence of autocorrelation in data (i.e., the violation of the assumption of exchangeability of observations) distorted the false alarm rates. Table 2 shows that positive serial dependence can lead to underestimation or overestimation of Type I error rates in comparison to independent data series, according to the data division. For an applied researcher, this would suppose increased probability of omitting an effective intervention or of a false alarm, respectively. Nonetheless, the effect of autocorrelation was only slight for the random intervention points ranging from 9 to 23, for which the randomization test is statistically valid. In the case of negative serial dependence (see Table 3), the results were similar to those found for positive autocorrelated data series. It should be noted that if the empirical Type I error rate is estimated regardless of the random intervention points, its value practically matches .0476, which is the value expected for a discrete uniform distribution with a total of 21 possible values.

TABLES 2 AND 3 ABOUT HERE

Discussion

The results of the present research suggest that applied researchers should be cautious when using the random intervention point randomization test studied here. Psychologists ought to know that if the data division randomly chosen contains 7 or less measurement times in either of the phases, there is high risk of labeling an ineffective treatment as effective. Therefore, in order to enhance the accuracy of the decision making process, applied researchers should be cautious if the selected intervention point is not between the 9th and the 23rd observation. There are two reasons for accepting only integers in the interval 9-23. First, if α is set equal to .05, the statistical test is valid. Second, although the exchangeability assumption has been violated in several experimental conditions of the simulation study, the randomization test is relatively robust for the random intervention points between the 9th and 23rd measurement. Also, note that this randomization test has zero power at $\alpha = .05$ if p_0 equals 6, 7, 8, 24, 25, or 26. If the random intervention point was equal to one of those values, statistical decision-making process should not be conducted and only descriptive statistical analysis should be carried out.

The rationale of the abovementioned recommendations can be found in the shape of the randomization distribution, which is used to obtain the *p*-value of the observed test statistic. It is often supposed that the statistic of interest follows a discrete uniform distribution when randomization tests are used to analyze the data resulting from single-case experiments. For example, if the number of possible random intervention points in an AB single-case design is equal to *q*, it is generally assumed that the minimal significance value equals 1/q (Edgington & Onghena, 2007). The present simulation study showed that this assumption is not met if data-division-specific randomization

distributions are taken into account for obtaining statistical significance. It would be suitable if the standard errors of the statistic were identical for each intervention point, but this does not hold for all random intervention points. The results of the present simulation suggest that, under the null hypothesis and for independent series, the minimal significance value does not equal 1/q if data-division-specific randomization distributions are considered. In other words, the shape of the distribution of the statistic depends upon the random intervention point being chosen, as the variance and kurtosis values showed. All data-division-specific randomization distributions were symmetrical and the mathematical expectancy equals the mean rank, the kurtosis values depending upon the random intervention point. That is, the randomization distribution of the statistic was conditioned to the random intervention point.

The question remains of why the data-division-specific randomization distribution does not, in general, follow a discrete uniform distribution in the randomization test studied. Suppose that the random intervention point was chosen and the outcome was computed. It should be noted that in most cases the data-division-specific randomization distribution of the statistic will be generated by bipartitions of data that vary in size. Therefore, data-division-specific randomization distributions would be composed of mixing phase lengths, and the standard errors of the statistic would be different for distinct permutations. Thus, given that the null hypothesis is true, large departures of the statistic value from zero are likely to occur in permutations based on clearly unequal group sizes. The present simulation has verified that the variance of the rank associated with the statistic value was larger in clearly unequal bipartition sizes than in approximately equal bipartition sizes. Although the data-division-specific randomization distributions are symmetrical, the mass moved from the center of the distribution to the tails as the bipartition of data were more unequal. If the common

randomization distribution is considered, the results concur with those corresponding to other simulation studies in which the common randomization distribution is analyzed instead of data-division-specific randomization distributions (Ferron & Ware, 1995). The common randomization distribution suppresses the marked deviations from the discrete uniform distribution that can be clearly identified in data-division-specific randomization distributions. The main reason for this fact is the differential kurtosis in data-division-specific randomization distributions. If the random starting point divides data into two markedly different series lengths, the distribution of the statistic becomes more platykurtic than the discrete uniform distribution. When the phase lengths are approximately equal, the data-division-specific randomization distribution is less platykurtic than the discrete uniform distribution.

The conclusions of the present study are restricted by the experimental conditions explored and its generalization to another set is not suggested. An AB design composed of 30 observations was considered because 21 possible random intervention points are required to reach a statistical significance value less than or equal to .05 if the intervention point is constrained to ensure that there will be at least five observations in A and B phase.

Future research could be directed towards studying whether the present results can be verified for larger data series and to analyze the power of this randomization test. In any case, the present simulation suggests that data-division-specific randomization distributions should be analyzed when the validity and power of randomization tests are studied.

References

Backman, C. L., & Harris, S. R. (1999). Case studies, single-subject research, and N of 1 randomized trials: Comparisons and contrasts. *American Journal of Psychical Medicine & Rehabilitation, 78,* 170-176.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40,* 467-478.

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

Edgington, E. S. (1967). Statistical inference from N = 1 experiments. *The Journal of Psychology, 65*, 195−199.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *The Journal of Psychology*, *90*, 57−68.

Edgington, E. S. (1980a). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 3*, 235−251.

Edgington, E. S. (1980b). 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., 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 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 the 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 applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107−118.

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*, 271−283.

Kazdin, A. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*, 629-642.

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 of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341−351.

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

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics*. New York: John Wiley & Sons.

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., 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.

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.

TABLES

TABLE 1. Estimated mean, variance, skewness, and kurtosis for the data-division-specific randomization distribution as a function of the randomly chosen intervention point. Statistics were calculated for the rank associated with the test statistic for the data at hand. Empirical Type I error rates for the data-division-specific randomization distribution are provided for both extreme ranks (Rank 1 and Rank 21), and their averages. When data are averaged for all random intervention points, the results correspond to the common randomization distribution. Data series were simulated with 30 observations, a minimum of 5 observations per phase, and $\varphi_1 = .0$.

Intervention point	Test statistic	Mean	Variance	Skewness	Kurtosis	Rank 1	Rank 21	(Rank 1+Rank 21) /2
6	1	11.000	55.574	.000	−1.551	.1436	.1449	.1443
	2	11.000	49.205	.001	−1.443	.1085	.1110	.1098
7	1	11.008	49.263	.000	−1.507	.0726	.0744	.0735
	2	10.998	45.479	.003	−1.427	.0647	.0664	.0656
8	1	11.005	43.605	.001	−1.441	.0515	.0533	.0524
	2	11.009	41.796	.000	−1.381	.0524	.0539	.0532
9	1	10.996	39.286	−.001	−1.368	.0416	.0423	.0420
	2	11.004	38.829	.001	−1.330	.0443	.0450	.0447
10	1	10.995	35.756	−.001	−1.276	.0365	.0343	.0354
	2	10.987	36.341	.001	−1.262	.0410	.0397	.0404
11	1	10.992	32.966	−.002	−1.173	.0341	.0319	.0330
	2	10.989	34.310	.000	−1.185	.0391	.0378	.0385
12	1	10.989	30.648	−.001	−1.076	.0285	.0286	.0286
	2	10.982	32.526	−.002	−1.113	.0350	.0343	.0347
13	1	11.009	28.939	−.005	−.969	.0283	.0269	.0276
	2	11.007	31.184	−.003	−1.035	.0344	.0336	.0340
14	1	11.017	28.171	−.003	−.907	.0257	.0270	.0264
	2	11.018	30.607	−.003	−.987	.0327	.0345	.0336
15	1	11.013	27.282	.007	−.838	.0258	.0264	.0261
	2	11.014	29.809	.006	−.932	.0324	.0334	.0329
16	1	11.022	27.099	−.001	−.822	.0261	.0267	.0264
	2	11.021	29.730	−.003	−.927	.0324	.0330	.0327
17	1	10.995	27.456	−.001	−.858	.0256	.0260	.0258
	2	10.997	30.030	−.001	−.951	.0328	.0320	.0324
18	1	10.993	28.023	.002	−.909	.0257	.0254	.0256
	2	10.933	30.513	.007	−.993	.0320	.0329	.0325
19	1	11.004	28.957	.001	−.973	.0277	.0263	.0270
	2	11.004	31.137	.006	−1.037	.0329	.0325	.0327
20	1	11.012	30.691	.004	−1.073	.0294	.0297	.0296
	2	11.008	32.547	.004	−1.109	.0355	.0354	.0355
21	1	10.995	32.989	.002	−1.181	.0319	.0319	.0319
	2	10.996	34.216	.002	−1.191	.0370	.0371	.0371
22	1	10.984	35.734	.005	−1.277	.0358	.0367	.0363
	2	10.987	36.280	.004	−1.264	.0403	.0404	.0404
23	1	10.985	39.140	.000	−1.358	.0439	.0415	.0427
	2	10.991	38.634	−.002	−1.321	.0462	.0438	.0450
24	1	10.965	43.523	.011	−1.440	.0520	.0512	.0516
	2	10.978	41.724	.008	−1.383	.0510	.0514	.0512
25	1	11.011	48.965	−.001	−1.506	.0708	.0717	.0713
	2	11.007	45.220	−.001	−1.423	.0641	.0657	.0649
26	1	11.004	55.729	.000	−1.553	.1435	.1444	.1440
	2	11.014	49.266	−.002	−1.444	.1100	.1103	.1102
Mean	1	11.000	36.657	.001	−1.193	.0476	.0477	.0476
	2	11.000	36.637	.001	−1.197	.0476	.0478	.0477
Standard Deviation	1	.013	9.181	.003	.250	.0339	.0344	.0342
	2	.012	6.353	.003	.182	.0222	.0227	.0224

TABLE 2. One-sided empirical Type I error rates as a function of the random intervention point. When data are averaged for all random intervention points, the results correspond to the common randomization distribution. Data series were simulated with 30 observations, a minimum of 5 observations per phase, and several positive values for the first order autoregressive parameter.

		φ = .3		φ = .6		φ = .9	
Intervention point	Test statistic	Rank 1	Rank 21	Rank 1	Rank 21	Rank 1	Rank 21
6	1	.1569	.1563	.1735	.1754	.2141	.2126
	2	.1218	.1220	.1374	.1386	.1730	.1739
7	1	.0627	.0641	.0555	.0565	.0499	.0476
	2	.0591	.0601	.0539	.0525	.0457	.0449
8	1	.0478	.0491	.0428	.0435	.0381	.0385
	2	.0477	.0472	.0430	.0450	.0382	.0376
9	1	.0406	.0409	.0362	.0382	.0304	.0336
	2	.0445	.0440	.0400	.0402	.0347	.0359
10	1	.0355	.0342	.0332	.0344	.0286	.0298
	2	.0399	.0391	.0370	.0376	.0328	.0352
11	1	.0324	.0325	.0310	.0312	.0272	.0277
	2	.0365	.0377	.0359	.0363	.0321	.0332
12	1	.0304	.0294	.0291	.0294	.0262	.0271
	2	.0355	.0349	.0335	.0350	.0318	.0322
13	1	.0281	.0287	.0283	.0281	.0253	.0249
	2	.0328	.0338	.0337	.0340	.0318	.0308
14	1	.0268	.0271	.0272	.0283	.0243	.0239
	2	.0332	.0336	.0334	.0333	.0299	.0299
15	1	.0267	.0262	.0266	.0277	.0244	.0245
	2	.0318	.0320	.0331	.0330	.0313	.0310
16	1	.0267	.0256	.0275	.0274	.0246	.0244
	2	.0324	.0319	.0327	.0331	.0314	.0317
17	1	.0254	.0265	.0264	.0264	.0233	.0249
	2	.0323	.0325	.0323	.0313	.0297	.0314
18	1	.0267	.0262	.0267	.0269	.0226	.0254
	2	.0326	.0322	.0320	.0316	.0295	.0314
19	1	.0283	.0276	.0282	.0285	.0245	.0245
	2	.0346	.0333	.0330	.0340	.0309	.0306
20	1	.0290	.0310	.0276	.0300	.0256	.0265
	2	.0349	.0354	.0334	.0352	.0314	.0318
21	1	.0330	.0317	.0316	.0320	.0289	.0276
	2	.0385	.0359	.0363	.0366	.0321	.0323
22	1	.0359	.0341	.0339	.0343	.0297	.0288
	2	.0400	.0389	.0386	.0381	.0342	.0331
23	1	.0405	.0397	.0371	.0358	.0319	.0302
	2	.0430	.0423	.0408	.0404	.0345	.0348
24	1	.0473	.0475	.0438	.0426	.0352	.0385
	2	.0476	.0488	.0456	.0441	.0387	.0390
25	1	.0626	.0631	.0536	.0545	.0479	.0501
	2	.0574	.0576	.0524	.0533	.0452	.0447
26	1	.1566	.1560	.1741	.1749	.2133	.2135
	2	.1217	.1242	.1382	.1383	.1733	.1742
Mean	1	.0476	.0475	.0473	.0479	.0474	.0478
	2	.0475	.0475	.0474	.0477	.0472	.0476
Standard	1	.0370	.0370	.0419	.0421	.0544	.0541
Deviation	2	.0253	.0258	.0300	.0300	.0411	.0412

TABLE 3. One-sided empirical Type I error rates as a function of the random intervention point. When data are averaged for all random intervention points, the results correspond to the common randomization distribution. Data series were simulated with 30 observations, a minimum of 5 observations per phase, and several negative values for the first order autoregressive parameter.

		φ = −.3		φ = −.6		φ = −.9	
Intervention point	Test statistic	Rank 1	Rank 21	Rank 1	Rank 21	Rank 1	Rank 21
6	1	.1363	.1357	.1334	.1326	.1777	.1796
	2	.1026	.1019	.0992	.0992	.1311	.1343
7	1	.0812	.0820	.0907	.0900	.0863	.0870
	2	.0704	.0710	.0760	.0755	.0699	.0706
8	1	.0544	.0532	.0546	.0557	.0465	.0443
	2	.0536	.0529	.0535	.0548	.0513	.0506
9	1	.0436	.0441	.0442	.0425	.0418	.0410
	2	.0458	.0471	.0472	.0465	.0439	.0450
10	1	.0358	.0355	.0358	.0353	.0270	.0265
	2	.0409	.0404	.0412	.0408	.0350	.0349
11	1	.0324	.0323	.0300	.0311	.0299	.0285
	2	.0381	.0384	.0371	.0384	.0362	.0349
12	1	.0293	.0284	.0279	.0268	.0199	.0200
	2	.0356	.0343	.0349	.0342	.0277	.0289
13	1	.0271	.0269	.0265	.0254	.0240	.0239
	2	.0336	.0343	.0346	.0327	.0320	.0324
14	1	.0243	.0249	.0223	.0245	.0170	.0170
	2	.0321	.0313	.0303	.0317	.0251	.0270
15	1	.0243	.0248	.0226	.0239	.0222	.0223
	2	.0318	.0318	.0307	.0316	.0305	.0308
16	1	.0237	.0247	.0229	.0222	.0168	.0161
	2	.0312	.0319	.0308	.0304	.0249	.0256
17	1	.0252	.0256	.0227	.0230	.0225	.0226
	2	.0326	.0334	.0308	.0316	.0308	.0303
18	1	.0246	.0237	.0233	.0244	.0184	.0161
	2	.0325	.0310	.0314	.0315	.0269	.0251
19	1	.0271	.0264	.0247	.0246	.0244	.0246
	2	.0340	.0336	.0327	.0323	.0319	.0325
20	1	.0282	.0285	.0268	.0286	.0216	.0199
	2	.0351	.0352	.0340	.0354	.0293	.0273
21	1	.0325	.0317	.0307	.0297	.0288	.0286
	2	.0387	.0371	.0364	.0362	.0351	.0345
22	1	.0365	.0367	.0359	.0354	.0265	.0260
	2	.0410	.0407	.0414	.0410	.0352	.0360
23	1	.0432	.0438	.0438	.0431	.0417	.0424
	2	.0458	.0458	.0472	.0467	.0437	.0446
24	1	.0527	.0542	.0545	.0535	.0458	.0457
	2	.0525	.0525	.0532	.0520	.0512	.0512
25	1	.0789	.0808	.0919	.0900	.0863	.0862
	2	.0690	.0700	.0766	.0757	.0708	.0713
26	1	.1330	.1348	.1332	.1342	.1783	.1781
	2	.1001	.1012	.0993	.0985	.1320	.1345
Mean	1	.0473	.0476	.0475	.0475	.0478	.0474
	2	.0475	.0474	.0478	.0475	.0474	.0477
Standard	1	.0326	.0329	.0341	.0338	.0464	.0468
Deviation	2	.0207	.0209	.0211	.0210	.0301	.0309

FIGURES

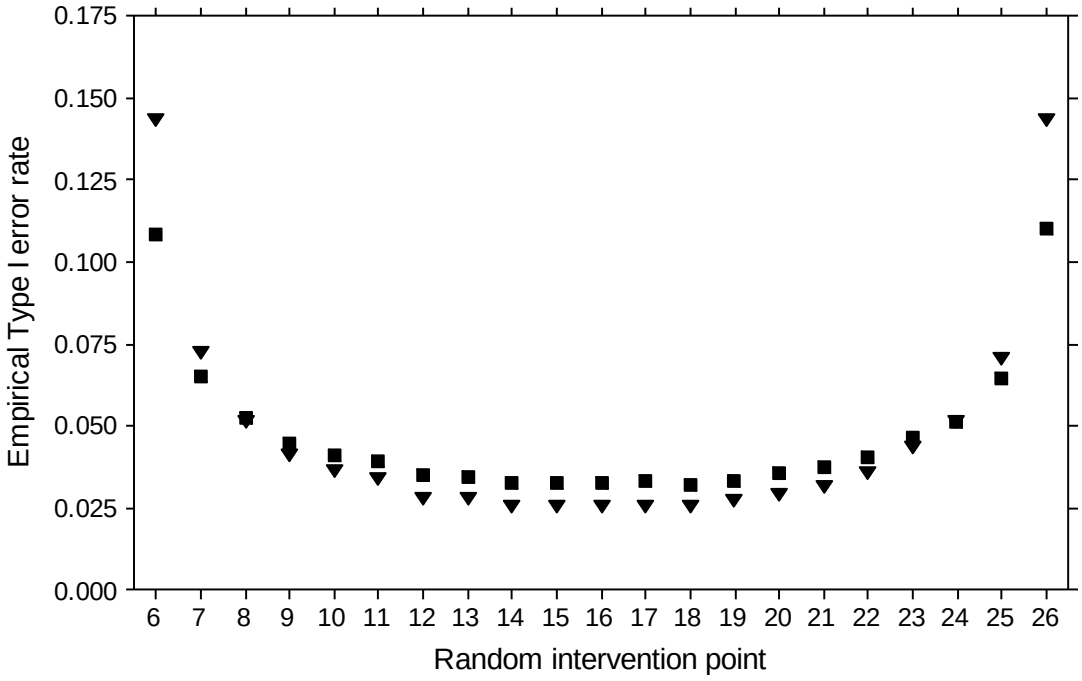

FIGURE 1. Empirical Type I error rates for independent data series as a function of the random intervention point and the test statistic used. The proportion values correspond to the most extreme ranks in the data-division-specific randomization distributions.

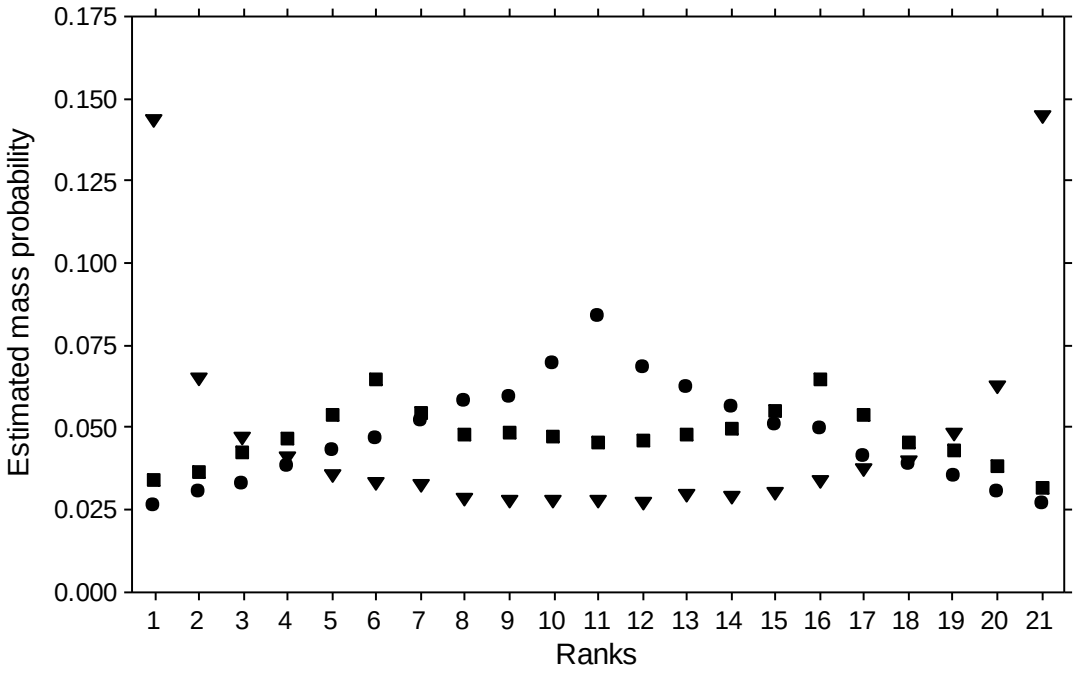

FIGURE 2. Estimated mass probability functions for three data-division-specific randomization distributions (random intervention points $p_o = 6$, $p_o = 11$, and $p_o = 16$) for *Statistic 1* and nonautocorrelated data series.

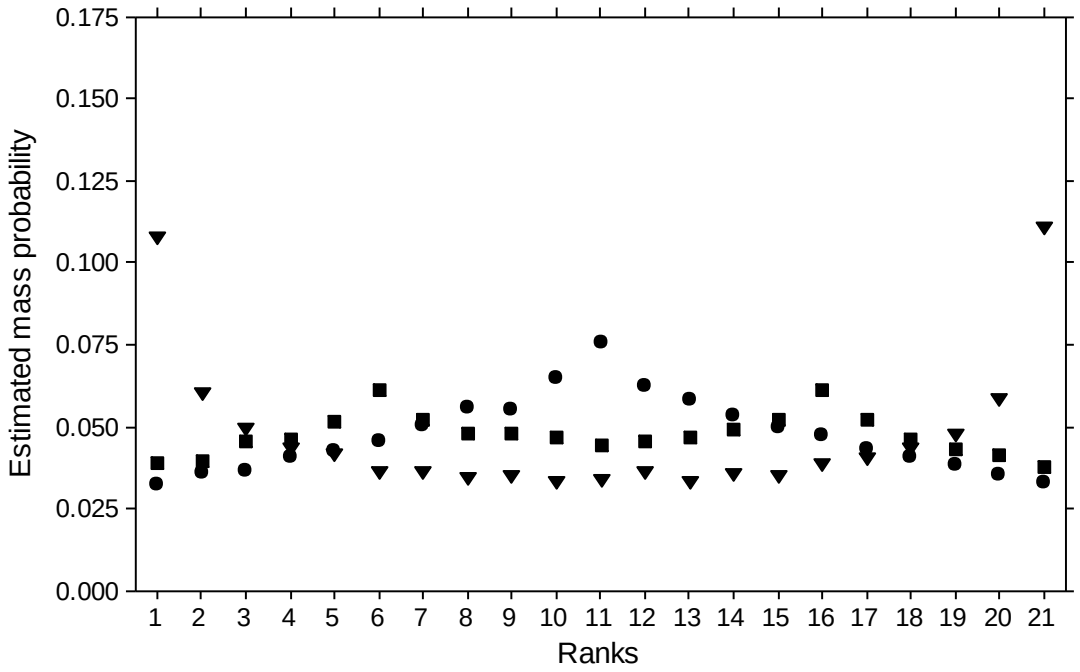

FIGURE 3. Estimated mass probability functions for three data-division-specific randomization distributions (random intervention points $p_o = 6$, $p_o = 11$, and $p_o = 16$) for *Statistic 2* and nonautocorrelated data series.