Psicothema 2008. Vol. 20, nº 2, pp. 297-303
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Randomization tests for ABAB designs: Comparing-data-division-specific and common distributions

Rumen Manolov and Antonio Solanas
Universidad de Barcelona

Monte Carlo simulations were used to generate data for ABAB designs of different lengths. The points of change in phase are randomly determined before gathering behaviour measurements, which allows the use of a randomization test as an analytic technique. Data simulation and analysis can be based either on data-division-specific or on common distributions. Following one method or another affects the results obtained after the randomization test has been applied. Therefore, the goal of the study was to examine these effects in more detail. The discrepancies in these approaches are obvious when data with zero treatment effect are considered and such approaches have implications for statistical power studies. Data-division-specific distributions provide more detailed information about the performance of the statistical technique.

Pruebas de aleatorización para diseños ABAB: comparación entre distribuciones específicas y comunes. A través de simulación Monte Carlo se generaron datos para diseños ABAB de diferente longitud. La prueba de aleatorización estudiada, como técnica analítica, requiere que los puntos de cambio de fase se determinen de manera aleatoria antes de registrar la conducta de interés. La simulación y el análisis de los datos se pueden fundamentar en una distribución de aleatorización común o en distribuciones específicas para cada división de datos. El hecho de seguir un método u otro afecta a los resultados obtenidos tras la aplicación de la prueba de aleatorización, por lo que el objetivo del estudio fue profundizar en estos efectos. En los resultados se hacen evidentes las discrepancias entre las dos aproximaciones en ausencia de efecto del tratamiento y éstas tienen implicaciones para el estudio de la potencia estadística. Las distribuciones específicas proporcionan una información más detallada sobre las propiedades estadísticas de la técnica analítica.

N= 1 designs are frequent and sometimes the only possible in applied research in cognitive, clinical or educational psychology (e.g., Crane, 1985; Martínez, Ortiz, & González, 2007; Olivencia & Cangas, 2005; Onghena & Edgington, 2005; Rabin, 1981; Tervo, Estrem, Bryson-Brockman, & Symons, 2003). The development of appropriate single-case (i.e., repeated measures) analytical techniques requires addressing the question of autocorrelation (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen & Ary, 1987). Mixed models are a viable option especially when multiple dependent variables are measured (Vallejo & Lozano, 2006). Randomization tests constitute another proposal (Edgington, 1995) and require fewer assumptions than classical parametric tests. Unlike visual inspection, formal decision rules are available for randomization tests through statistical significance values. Moreover, this procedure does not require as much observation points and analysts' expertise as interrupted time series analysis. Nevertheless, the statistical properties of randomization test have to be investigated in the presence of different degrees of serial dependency in order to establish the validity of the technique as defined by Edgington (1980b) and Hayes (1996).

Although randomization tests are not so widely known and applied as ANOVA or ARIMA, a considerable amount of research has already been produced. Among the designs studied can be found AB (Ferron & Ware, 1995), ABAB (Ferron, Foster-Johnson, & Kromrey, 2003; Ferron & Ware, 1995; Levin, Marascuilo, & Hubert, 1978), ABABAB (Ferron & Onghena, 1996), multiple baseline (Ferron & Sentovich, 2002; Wampold & Worsham, 1986), responsive designs (Ferron & Ware, 1994), simultaneous treatment designs (Kratochwill & Levin, 1980), restricted alternating treatments design (Onghena & Edgington, 1994), replicated AB (Lall & Levin, 2004; Levin & Wampold, 1999) and replicated ABAB designs (Marascuilo & Busk, 1988). Most of these investigations suggested that Type I error rates are generally controlled in presence of autocorrelation. Power seems adequate for some designs (Ferron & Onghena, 1996) and insufficient for others (Ferron & Ware, 1995).

The application of randomization tests requires the randomization of some aspect of the design prior to data collection. In N= 1 studies, one possibility is to randomly assign the measurement times to the conditions (baseline and treatment). The procedure followed in the present study refers to phase

Fecha recepción: 20-7-07 • Fecha aceptación: 11-10-07
Correspondencia: Rumen Manolov
Facultad de Psicología
Universidad de Barcelona
08035 Barcelona (Spain)
E-mail: rrumenov13@ub.edu

designs (i.e., random determination of phase change) rather than to alternation designs, using the terms of Onghena & Edgington (2005). Another feature of the technique is that the reference set to which the value of the test statistic is compared is not a «sampling» but a «randomization» distribution as it arises from the permutations of data divisions. Hence, there is an implicit conditioning inherent to randomization tests as the statistical significance depends on the data at hand. One way of analyzing the statistical properties of randomization tests is based on the idea that the randomization distribution and statistical significance do not depend on the specific data division chosen. We refer to this approach as the «common distribution» and it is the one followed by the studies cited in the previous paragraph. This method estimates Type I and Type II error rates as a function of, for example, the degree of serial dependence (φ) and the series length (n). It assumes that the randomization test has similar (if not equivalent) performance for each data division, that is to say, regardless of the length of the individual phases (e.g., n_{A1}, n_{B1}, n_{A2}, and n_{B2} for an ABAB design).

Another way of studying randomization tests involves what we denominate «data-division-specific distributions». According to this approach the randomization distribution is different for each data division. As a result, the same value of the test statistic obtained from distinct data divisions may have a different position in the reference set and, hence, the significance level is not the same. Recent investigations have already tried to show the discrepancies in the results between the two methods for a variety of designs, for instance, AB (Manolov & Solanas, 2007) and ABABAB (Sierra, Solanas, & Quera, 2005). However, straightforward comparisons between studies are not always suitable due to imperfect correspondence between design lengths, autocorrelation levels and effect sizes studied. Other procedural divergences cannot be ruled out unless the same authors conduct an investigation including both approaches prior to comparing them.

Thus, the objective of the present study is to apply randomization tests to single-case data generated by means of Monte Carlo simulations. The assessment of the statistical properties of the data analysis technique will be carried out using data-division-specific as well as common distributions in order to show the incongruity of the conclusions that can be made according to which method is employed as a fundament.

Method

Selection of designs

The random assignment procedure for the ABAB designs studied is based on Onghena's (1992) proposal for randomly selecting the three points of change in phase. The restriction of a minimal phase (i.e., n_{A1}, n_{B1}, n_{A2}, and n_{B2}) length is applied to both approaches and it implies that the points of change are not independent. To make the selection each data division equally probable the random assignment is performed choosing from a list of all admissible «triplets» (data divisions are called «triplets» and are symbolized by «b1.a2.b2» – the first points of the last three phases in the ABAB design). The following design lengths were studied:

a) *n= 20*. A minimum of three data points per phase was used as it allows choosing out of a total of 165 different data divisions (following the formula presented in Onghena, 1992) and permits evaluating potential tendencies in the data belonging to each phase. Although more observations per condition would lead to a more precise estimate on the behaviour of the experimental unit in each phase, the length of the design is inversely related to its applicability in applied settings.

b) *n= 25*. In this case the minimal phase length was 4. The total number of different triplets is 220.

c) *n= 30*. This design length (equal to the one used by Ferron, Foster-Johnson, & Kromrey, 2003) permits, following Edgington's (1980a) recommendations, establishing a minimum of five measurement times per phase. As a consequence, there are 286 different triplets.

Data generation

Data was generated applying an equation commonly used (e.g., Ferron, Foster-Johnson, & Kromrey, 2003; Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Matyas & Greenwood, 1990) in research focused on similar topics: $y_t = \varphi_1 y_{t-1} + \varepsilon_t + d$, where:

y_t: datum obtained at measurement time t;
y_{t-1}: previous datum;
φ_1: value of the lag-one autocorrelation coefficient;
ε_t: independent error;
d: effect size.

We only focus on this first-order autoregressive model as it is the most common one in investigations centred on the same topic and due to the fact that we aim to show the difference in Type I and Type II error estimates between the studies referenced above and the present one. By means of the FORTRAN 90 programming language various programs were constructed in order to: a) make calls to the external subroutines *nag_rand_seed_set* and *nag_rand_normal* of NAG fl90 mathematical-statistical libraries for the generation of an error term (ε_t) according to a N(0,1); b) apply the formula previously specified; c) calculate the values and ranks of the test statistics.

Different levels of serial dependency (φ_1) were represented by values (–0.3, 0.0, 0.3 y 0.6) that assumingly represent adequately the characteristics of real data and, therefore, have been frequently used in simulation studies (Ferron & Onghena, 1996; Ferron & Ware, 1995; Greenwood & Matyas, 1990).

Effect size (d) was calculated as the difference between phase means divided by the standard deviation of the error term (ε_t), as defined in Ferron & Sentovich (2002). When the Type I error rates were the centre of interest, d was set to zero. The statistical power of the randomization tests was estimated for various effect sizes: 0.20, 0.50, 0.80, 1.10, 1.40, 1.70, and 2.00, replicating values from previous studies (e.g., Ferron & Onghena, 1996; Ferron & Sentovich, 2002). Effect sizes were applied in a way to produce an immediate increment in behaviour as each of the B-phases starts, while also maintaining that change in level throughout the whole phase.

Simulation

The simulation consisted of the following steps: 1) selection of an admissible triplet; 2) systematic selection of each of the four

levels of autocorrelation studied; 3) systematic selection of each effect size; 4) generation of data according to the equation presented; 5) calculation of the test statistics for the actual data, obtaining the outcome; 6) calculation of the tests statistics for each possible (and not selected) triplet; 7) ranking of the outcome according to its position in the reference set; 8) calculation of the proportion of critical ranks (only for the power study).

Simulation: common distributions

Step 1 was iterated 100,000 times, which seems enough to allow a precise estimation of Type I and II error rates (Robey & Barcikowski, 1992), as the common distributions approach does not imply a separate estimation for each data division but only a global one for each level of autocorrelation. For the study of Type I error rates *d* was invariably zero, while steps 4 to 7 were repeated for the four degrees of serial dependence studied. Hence, there were 100,000 * 4= 400,000 designs generated for each design length.

For the study of Type II error rates, steps 4 to 7 were repeated for each combination of autocorrelation level and effect size producing a total of 2,800,000 samples for each design length (*n*).

Simulation: data-division-specific distributions

In order to estimate Type I error rates for all triplets, step 1 involved the systematic selection of the data divisions and, therefore, was repeated 165 or 220 or 286 times according to design's length. Step 2 was repeated 4 times and steps 4 to 7 – 100,000 times defining a total number of iterations of 66,000,000 for *n*= 20; 88,000,000 for *n*= 25; and 114,400,000 for *n*= 30. Setting an equal number of iterations for estimating empirical error rates in both methods (common and data-division-specific) guarantees a proper comparison between each pair of experimental conditions.

In the Type II error rates study were included only the triplets for which the randomization test was robust against the violation of the independence assumption. The different combinations of iteration, triplet, values of φ_1 and value of *d* determine the number of samples simulated: a) 100,000 * 33 * 4 * 7= 92,400,000 for *n*= 20; b) 100,000 * 40 * 4 * 7= 112,000,000 for *n*= 25; c) 100,000 * 57 * 4 * 7= 159,600,000 for *n*= 30.

Analysis

We contrasted the following null hypothesis at 5% alpha:

H_0: $\bar{X}_A \geq \bar{X}_B$

Two test statistics were used:

a) Mean difference (hereinafter, MD): $\bar{X}_B - \bar{X}_A$.

b) Student's *t* (hereinafter, ST): $\left(\bar{X}_B - \bar{X}_A\right) / \sqrt{s^2/n_B + s^2/n_A}$.

The pooled variance estimate is used, as the process has been generated with a common variance for all phases.

Each test statistic is calculated on the actual data (i.e., the triplet selected) and after its value is located in the randomization distributions, the rank assignation takes place. The proportion of each rank is then calculated and the empirical distribution of ranks is obtained. When considering the common method, there is one ranks' distribution for each combination of degree of serial dependence and test statistic (reaching a total of 8), while the data-division-specific method implied that each data division had its own eight distributions of ranks.

In the absence of treatment effect and for uncorrelated data it was important to identify how many ranks enter the 5% barrier defined by the significance level chosen and can be regarded as «critical» for rejecting the null hypothesis. After that, robustness intervals following Bradley's stringent criterion (1978, cited in Robey & Barcikowski, 1992), $\alpha \pm 0.1\ \alpha$, were constructed around the cumulative proportions of ranks identified. The relative frequencies of the same number of ranks for $\varphi_1 = -0.3$, 0.3, and 0.6 were compared to the robustness intervals. In case there was no underestimation or overestimation in presence of autocorrelation, the randomization test was judged to be robust. The decision regarding robustness was made several times: a) following the common method – twice (in accordance with the number of test statistics) for each design length; b) following the data-division-specific method: 2 * 165 times for *n*= 20, 2 * 220 times for *n*= 25, and 2 * 286 times for *n*= 30.

For the cases in which the randomization test was robust, the corresponding number of critical ranks was used as cut-off points for the power analysis. The proportion of occasions in which the outcome has been assigned one of the critical ranks was the value of interest. This procedure resulted in one power estimate for each combination of φ_1, *d* and test statistic for common distributions, while for data-division-specific distributions there were different estimates for the distinct triplets.

Results

In this section only part of the data will be presented in tabular or graphical format. Full tables are available from the authors upon request.

Triplet selection

With data-division-specific distributions, each of the specific data divisions appears exactly 100,000 times in a systematic manner. The common simulation involves the random selection of a triplet in each of the 100,000 iterations and calculations showed that it resulted in each data divisions appearing approximately the same number of times. However, we should emphasize two questions. First, the common distributions approach makes no triplet-specific estimations but only global ones. Second, it is necessary to distinguish between the data-division-specific method of performing a simulation study on a randomization test (which involves non-random determination of the triplet) and the procedure that should be followed in an applied setting where the application of a randomization test requires the random selection of a data division. The data-division-specific approach is not supposed to encourage non-random selection of points of change in phase, but rather to acquaint applied researchers with most appropriate cut-off ranks for null-hypothesis rejection for each triplet.

Distribution of ranks

Simulations based on common distributions suggest that the distribution of ranks for the studied experimental conditions is a

uniform one irrespective of the autocorrelation level (see figures 1, 2, and 3). When we consider data-division-specific distributions, however, it is obvious that the uniform distribution is the less frequent one even for independent data series. According to the triplet chosen, the distribution of ranks may actually be uniform (figure 4), but it also may be U-shaped (figure 5), approximately triangular (figure 6) or bimodal (figure 7). An important sequel of these distributional differences was the incongruity among the number of critical ranks. This finding was common to all design lengths studied and this is the reason for suppressing redundant graphical representations.

Robustness

Another discrepancy between common and data-division-specific distributions was made evident when making the robustness decisions. When the former method was used, the randomization test was declared unaffected by autocorrelation (in general) for all *n*. In contrast, the latter approach indicated that the statistical technique was robust only for certain data divisions. The common approach yielded the following number of critical extreme ranks: 8 for *n*= 20, 11 for *n*= 25, and 14 for *n*= 30. There proved to be no difference between the test statistics used. The data-division-specific approach resulted in a different number of critical ranks for the distinct triplets and the two test statistics constituted another source of variation: a) *n*= 20: 33 robust triplets out of 165; critical ranks varied from 2 to 12; b) *n*= 25: 40 robust

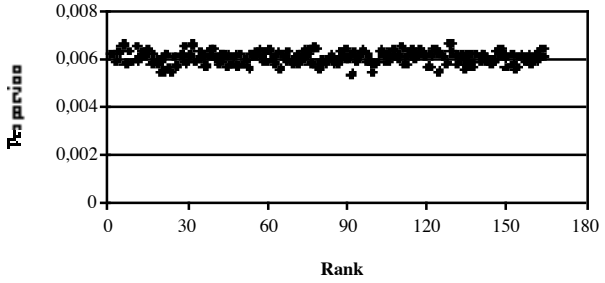

Figure 1. *Ranks distribution obtained using common distributions; n= 20;*

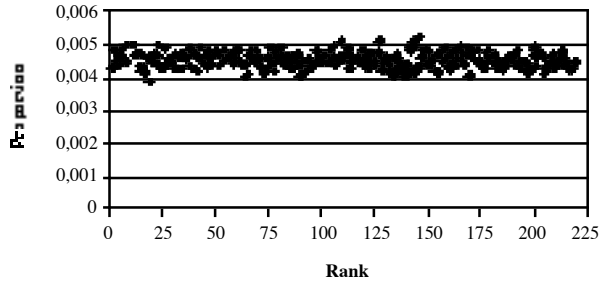

Figure 2. *Ranks distribution obtained using common distributions; n= 25;*

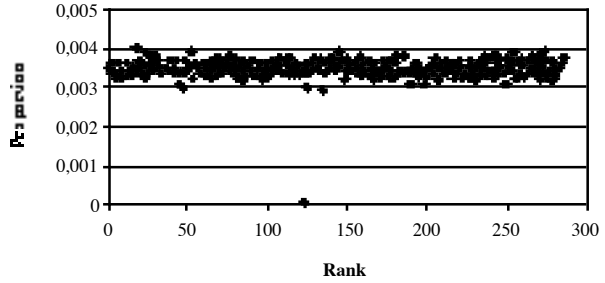

Figure 3. *Ranks distribution obtained using common distributions; n= 30;*

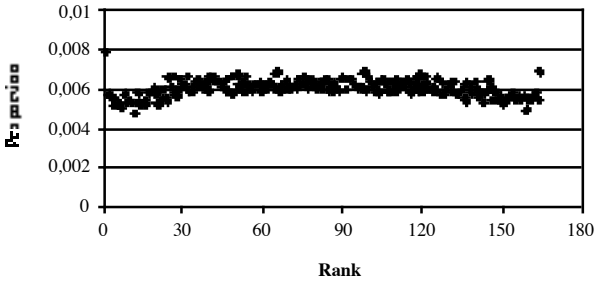

Figure 4. *Ranks distribution obtained using data-division-specific distri -*

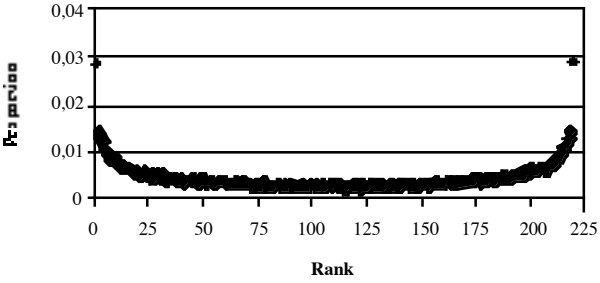

Figure 5. *Ranks distribution obtained using data-division-specific distri -*

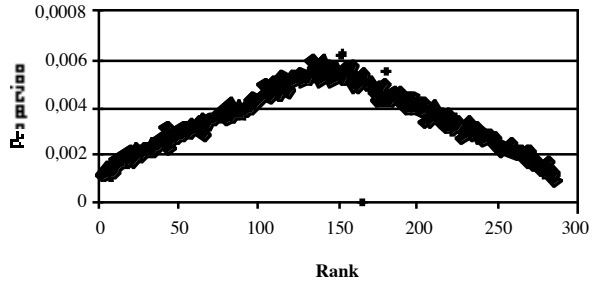

Figure 6. *Ranks distribution obtained using data-division-specific distri -*

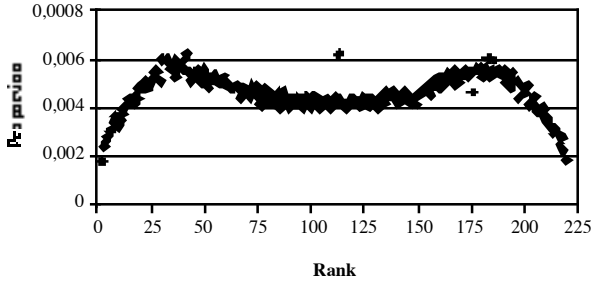

Figure 7. *Ranks distribution obtained using data-division-specific distri -*

triplets out of 220; critical ranks varied from 3 to 16; c) n= 30: 57 robust triplets out of 286; critical ranks varied from 4 to 22.

Power

Sensitivity to treatment effects assessed by both methods showed less incongruity. The power estimates obtained for equal design lengths were quite similar (as it can be seen from tables 1 to 6), but we have to keep in mind that there is no specific information about a considerable fraction of triplets, as they proved not to control the Type I error rate. Generally, the randomization test has adequate ($1-\beta \geq 0.80$; in Cohen's (1992) terms) power for d= 2.0, and in some cases even for d= 1.7, while φ_1= 0.6 is associated with insensitivity.

Table 1
Power estimates for n= 20 using the common approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0848	0.1700	0.2900	0.4381	0.5846	0.7173	0.8203
	ST	*0.0851*	*0.1712*	*0.2928*	*0.4420*	*0.5892*	*0.7224*	*0.8263*
0.0	MD	0.0866	0.1766	0.3064	0.4611	0.6177	0.7507	0.8510
	ST	*0.0868*	*0.1774*	*0.3093*	*0.4651*	*0.6231*	*0.7570*	*0.8577*
0.3	MD	0.0845	0.1686	0.2871	0.4273	0.5715	0.6964	0.7942
	ST	*0.0853*	*0.1694*	*0.2895*	*0.4306*	*0.5755*	*0.7012*	*0.7986*
0.6	MD	0.0758	0.1399	0.2200	0.3006	0.3728	0.4314	0.4736
	ST	*0.0764*	*0.1399*	*0.2203*	*0.2997*	*0.3719*	*0.4292*	*0.4703*

Table 2
Power estimates for n= 25 using the common approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0916	0.1945	0.3281	0.4892	0.6442	0.7732	0.8644
	ST	*0.0934*	*0.1948*	*0.3298*	*0.4927*	*0.6477*	*0.7761*	*0.8686*
0.0	MD	0.0924	0.1967	0.342	0.5118	0.6733	0.8038	0.8928
	ST	*0.0926*	*0.1977*	*0.3442*	*0.5146*	*0.6768*	*0.8076*	*0.8973*
0.3	MD	0.0916	0.1883	0.3217	0.4780	0.6271	0.7470	0.8391
	ST	*0.0918*	*0.1890*	*0.3235*	*0.4820*	*0.6311*	*0.7512*	*0.8441*
0.6	MD	0.0845	0.1537	0.2413	0.3338	0.4048	0.4628	0.5030
	ST	*0.0843*	*0.1547*	*0.2417*	*0.3334*	*0.4041*	*0.4606*	*0.4991*

Table 3
Power estimates for n= 30 using the common approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0924	0.2002	0.3520	0.5190	0.6777	0.8063	0.8929
	ST	*0.0927*	*0.201*	*0.3525*	*0.5220*	*0.6804*	*0.8096*	*0.8955*
0.0	MD	0.0925	0.2058	0.3691	0.5428	0.7063	0.8329	0.9152
	ST	*0.0930*	*0.2060*	*0.3700*	*0.5472*	*0.7115*	*0.8362*	*0.9180*
0.3	MD	0.0933	0.1963	0.3440	0.5102	0.6600	0.7821	0.8677
	ST	*0.0934*	*0.1973*	*0.3466*	*0.5141*	*0.6646*	*0.7870*	*0.8717*
0.6	MD	0.0861	0.1630	0.2584	0.3547	0.4278	0.4845	0.5273
	ST	*0.0861*	*0.1632*	*0.2595*	*0.3545*	*0.4276*	*0.4845*	*0.5248*

Table 4
Mean power estimates (based on 30 triplets for MD and 31 triplets for ST) for n= 20 using the data division specific approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0818	0.1664	0.2821	0.4275	0.5772	0.7205	0.8244
	ST	*0.0819*	*0.1669*	*0.2839*	*0.4303*	*0.5800*	*0.7236*	*0.8266*
0.0	MD	0.0839	0.1738	0.3050	0.4622	0.6212	0.7575	0.8586
	ST	*0.0838*	*0.1742*	*0.3062*	*0.4643*	*0.6233*	*0.7592*	*0.8596*
0.3	MD	0.0828	0.1668	0.2870	0.4310	0.5747	0.6824	0.7979
	ST	*0.0825*	*0.1671*	*0.2879*	*0.4308*	*0.5738*	*0.6984*	*0.7956*
0.6	MD	0.0754	0.1392	0.2191	0.3028	0.3783	0.4399	0.4875
	ST	*0.0753*	*0.1390*	*0.2183*	*0.3007*	*0.3746*	*0.4342*	*0.4807*

Table 5
Mean power estimates (based on 39 triplets for MD and 38 triplets for ST) for n= 25 using the data division specific approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0856	0.1769	0.3195	0.4812	0.6383	0.7692	0.8642
	ST	*0.0875*	*0.1851*	*0.3245*	*0.4872*	*0.6439*	*0.7737*	*0.8672*
0.0	MD	0.0876	0.1898	0.3367	0.5093	0.6724	0.8029	0.8930
	ST	*0.0900*	*0.1942*	*0.3431*	*0.5161*	*0.6781*	*0.8068*	*0.8953*
0.3	MD	0.0868	0.1830	0.3196	0.4767	0.6261	0.7487	0.8393
	ST	*0.0896*	*0.1881*	*0.3269*	*0.4841*	*0.6324*	*0.7532*	*0.8421*
0.6	MD	0.0798	0.1517	0.2423	0.3294	0.4057	0.4646	0.5085
	ST	*0.0829*	*0.1567*	*0.2463*	*0.3360*	*0.4124*	*0.4702*	*0.5150*

Table 6
Mean power estimates (based on 52 triplets for MD and 51 triplets for ST) for n= 30 using the data division specific approach

φ	test statistic	d= 0.2	d= 0.5	d= 0.8	d= 1.1	d= 1.4	d= 1.7	d= 2.0
-0.3	MD	0.0925	0.2014	0.3558	0.5297	0.6914	0.8165	0.9009
	ST	*0.0924*	*0.2011*	*0.3550*	*0.5287*	*0.6899*	*0.8029*	*0.8996*
0.0	MD	0.0923	0.2056	0.3543	0.5473	0.7094	0.8303	0.9068
	ST	*0.0940*	*0.2096*	*0.3745*	*0.5580*	*0.7233*	*0.8462*	*0.9241*
0.3	MD	0.0925	0.2015	0.3553	0.5248	0.6782	0.7681	0.8495
	ST	*0.0926*	*0.2031*	*0.3579*	*0.5283*	*0.6818*	*0.7866*	*0.8827*
0.6	MD	0.0851	0.1673	0.2666	0.3622	0.4405	0.4999	0.5431
	ST	*0.0858*	*0.1699*	*0.2714*	*0.3692*	*0.4489*	*0.5091*	*0.5527*

Discussion

The results reported by the present study show that Type I error rates vary across data divisions when the observations are independent (i.e., there is no assumption violated). This is due to the fact that the shape of the randomization distribution is not the same for all data divisions, contrary to what the common distribution approach assumes. If in the absence of serial dependence statistical decisions are made on the basis of the common distribution (e.g., using the 11 extreme ranks to reject the null hypothesis in a n= 25 design), the empirical Type I error rate

may diverge from the nominal one for each data division. If the randomization distribution for that specific data division is U-shaped (as it is for «5.13.19») then the cumulative proportion of the 11 extreme ranks would be greater than 0.05 and, therefore, the probability of committing a Type I error would be greater than 5%. If the randomization distribution is triangular (e.g., «5.14.20») or bimodal (e.g., «6.11.18»), then rejecting the null hypothesis when the outcome is assigned one of the 11 extreme ranks would render the test too conservative and less powerful. Basing the statistical decision on the common distribution would not be detrimental only when the randomization distribution is uniform for the specific data division, as would be the case for «5.16.21».

Another concern arises from the fact that while studies based on the common distribution suggest that randomization tests are not affected by autocorrelation, the data-division-specific approach shows that this is not always true. The present study shows that the effect of serial dependence is not the same for all data divisions and, therefore, robustness studies based on the common distribution assuming invariability of the effect provide insufficient information. The matching between nominal and empirical Type I error rates alleged by common distribution investigations can be explained by the fact that all data divisions are mixed, which leads to averaging the proportions of the ranks assigned to the outcome, and those ranks are the fundament of the Type I and Type II error estimates. Mixing triplets does not have any correspondence to reality as applied researchers make statistical decisions based on a specific data division in their particular case. By means of the data-division-specific approach we get to know which the number of critical ranks is for each specific triplet according to the shape of the randomization distribution for that triplet. Therefore, applied researchers still have to select randomly one of the triplets, but when analyzing data, investigations based on the data-division-specific approach will inform them about which are the most appropriate critical ranks for each data division and what is the probability of committing Type I and Type II errors using those ranks.

The similarity between the power estimates obtained via each of the two approaches is evident. Nevertheless, the consequence of using, for each data division, the same number of critical ranks (as suggested by the common approach) has to be emphasized. This would lead to a very powerful test for some data divisions (with greater probability for committing a Type I error) and to a more insensitive test for other data divisions (i.e, increased probability of Type II errors). Therefore, statistical decision making would become less reliable.

In this study only designs following the ABAB-structure were investigated in order to illustrate the disagreement between data-division-specific and common approaches. However, evidence from other studies (Manolov & Solanas, 2007) suggests that this finding is not exclusive to four-phase designs. A more general limitation applicable to randomization tests is the implicit assumption that the data series produced by a participant are the only ones possible. We deem that the measurements originated by an individual are just a random sample of all possible measurements that he or she could have generated, something that has no relation with random sampling of participants from a population.

We consider that, given the evidence presented here and taking into account the previous investigations mentioned, future simulation research on randomization tests should be made employing the data-division-specific approach. This approach is beyond doubt more time-consuming, but the efforts pay off as the more detailed information obtained is relevant for a more profound knowledge of the performance of the statistical technique studied. It has to be kept in mind that the statistical properties of randomization tests depend on the specific design structure (phase order) and data division (individual phase length) and we agree with Phillips (1983) that the influence of autocorrelation has to be studied for each design. Therefore, it is important to obtain information on the performance of randomization tests for the most frequently used designs in applied behavioural research in order to know for which data divisions the technique will be useful in establishing the existence of treatment effects.

Authors' note

This research was supported by the *Departament d'Educació i Universitats de la 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.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments and suggestions, which contributed to improving the manuscript.

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.

Crane, D.R. (1985). Single-case experimental designs in family therapy research: Limitations and considerations. *Family Process, 24*, 69-77.

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

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.

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. (1994). Using randomization tests with responsive single-case designs. *Behaviour Research and Therapy, 32*, 787-791.

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

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: ρ= 0. *Psychological Methods, 1*, 184-198.

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. (2007). *Validity and power of the random intervention point randomization test for an AB design*. Paper presented at the 10th European Congress of Psychology, Prague, Czech Republic.

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

Martínez, H., Ortiz, G., & González, A. (2007). Efectos diferencias de instrucciones y consecuencias en ejecuciones de discriminación condicional humana. *Psicothema, 19*, 14-22.

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.

Olivencia, J.J., & Cangas, A.J. (2005). Tratamiento psicológico del trastorno esquizotípico de la personalidad. Un estudio de caso. *Psicothema, 17*, 412-417.

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.

Onghena, P., & Edgington, E.S. (2005). Customization of pain treatments: Single-case design and analyses. *Clinical Journal of Pain, 21*, 56-68.

Phillips, J.P.N. (1983). Serially correlated errors in some single-subject designs. *British Journal of Mathematical and Statistical Psychology, 36*, 269-280.

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

Tervo, R.C., & Estrem, T.L., Bryson-Brockman, W., & Symons, F.J. (2003). Single-case experimental designs: Application in developmental-behavioral pediatrics. *Developmental and Behavioral Pediatrics, 24*, 438-448.

Vallejo, G., & Lozano, L.M. (2006). Modelos de análisis para los diseños multivariados de medidas repetidas. *Psicothema, 18*, 293-299.

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