DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 2

11-2014

Improved Randomization Tests for a Class of Single-Case Intervention Designs

Joel R. Levin
University of Arizona, jrlevin@u.arizona.edu

John M. Ferron
University of South Florida, ferron@usf.edu

Boris S. Gafurov
George Mason University, bgafurov@gmu.edu

Recommended Citation
Levin, Joel R.; Ferron, John M.; and Gafurov, Boris S. (2014) "Improved Randomization Tests for a Class of Single-Case Intervention Designs," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 2 , Article 2.
DOI: 10.22237/jmasm/1414814460
Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2

Improved Randomization Tests for a Class of Single-Case Intervention Designs

Erratum

In the original published version of this article, Panels B and C of Figure 15, p. 38, were reversed, and references to Figure 11 on p. 32 should have referred to Figure 12. This has been corrected.

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 2-52.

ISSN 1538 − 9472

Invited Article: Improved Randomization Tests for a Class of Single-Case Intervention Designs

Joel R. Levin
University of Arizona
Tucson, AZ

John M. Ferron
University of South Florida
Tampa, FL

Boris S. Gafurov
George Mason University
Fairfax, VA

Forty years ago, Eugene Edgington developed a single-case AB intervention design-and-analysis procedure based on a random determination of the point at which the B phase would start. In the present simulation studies encompassing a variety of AB-type contexts, it is demonstrated that by also randomizing the order in which the A and B phases are administered, a researcher can markedly increase the procedure's statistical power.

Keywords: Single-case intervention research, design and statistical analysis, randomization tests, statistical power, internal validity, scientific credibility

Introduction

Single-case designs that focus on behavioral and academic interventions are prevalent in a variety of clinical and educational fields (see, for example, Kratochwill & Levin, 2014). In contrast to conventional group intervention designs, single-case designs typically include only one or a few units (e.g., individuals, small groups, classrooms) to whom the intervention is administered. In addition, single-case intervention designs are intensive and implemented over longer periods of time, with more numerous assessments of the outcome measures (Horner & Odom, 2014; Kratochwill et al., 2010). Single-case intervention designs that currently incorporate formal criteria to enhance their scientific

Joel R. Levin is a Professor Emeritus in the Department of Educational Psychology. Email at jrlevin@u.arizona.edu. John M. Ferron is a Professor in the Department of Educational and Psychological Studies in the College of Education. Email at ferron@usf.edu. Boris S. Gafurov is an Assistant Professor in the College of Education and Human Development. Email at bgafurov@gmu.edu.

Erratum: *In the original published version of this article, Panels B and C of Figure 15 were reversed. This error is now corrected.*

credibility (Levin, 1994) include ABAB designs, alternating treatment designs, and multiple-baseline designs (Kratochwill et al., 2013).

As the methodological rigor of single-case intervention designs has evolved over the years (Kratochwill & Levin, 2010), so too have the formal statistical-analysis procedures that accompany them (see, for example, Kratochwill & Levin, 2014; and Manolov, Evans, Gast, & Perdices, 2014). Although various visual/graphical approaches remain an analytic staple of single-case data (e.g., Auerbach & Zeitlin, 2014; Kratochwill, Levin, Horner, & Swoboda, 2014; Parker, Vannest, & Davis, 2014), improved statistical methods have increasingly been considered as viable supplements to visual analysis. These improved statistical methods include econometric time-series analyses (e.g., McCleary & Welsh, 1992), adapted regression- and hierarchical linear modeling procedures (e.g., Maggin et al., 2011; Manolov & Solanas, 2013; Moeyaert, Ferron, Beretvas, Van den Noortgate, & Beretvas, 2014; Shadish, Kyse, & Rindskopf, 2013), and nonparametric permutation and randomization tests (e.g., Edgington & Onghena, 2007; Ferron & Levin, 2014; Heyvaert & Onghena, 2014). The last of these statistical approaches is the focus of the present study.

Overview of the Present Study

The motivation for single-case researchers to adopt a randomization test as one component of their analytic armament is that randomization tests provide strict control of the Type I error rate (i.e., the probability of concluding that phase-to-phase differences in level, trend, variability, etc. are present when those differences are simply chance fluctuations) as long as: (1) the design includes randomization; (2) the accompanying statistical test is conducted in a manner that is consistent with the design frame; and (3) the test statistic is chosen without knowledge of the results (Edgington, 1980; Ferron & Levin, 2014). In contrast, demonstration of Type I error control has been elusive in studies of visual analysis (e.g., Ferron & Jones, 2006; Fisch, 2001; Stocks & Williams, 1995). Moreover, with regression and hierarchical models, Type I error control hinges on a relatively strong set of assumptions (Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014). The modeling assumptions include: (1) the error distribution is correctly specified (e.g., normally distributed, homogeneous variances across phases, and a first-order autoregressive function); (2) the baseline trajectory is correctly specified; (3) the baseline trajectory can be extrapolated (i.e., had the intervention not been implemented, the baseline trajectory would have continued, implying that there were no confounding effects of external events on the time series); and (4) the treatment phase trajectory is correctly specified. Accordingly, a single-case researcher may plan a multicomponent analysis in which visual analysis serves as the primary analysis tool, a randomization test is employed to ensure that the Type I error rate is controlled, and a regression-based or hierarchical linear model is examined to summarize and estimate the size of the effect(s).

A concern with the addition of randomization tests to the analytic plan is that such tests require the researcher to introduce randomization into the design, and if the randomization is not carefully planned it can lead to a design that falls short of single-case design standards (e.g., Ferron & Levin, 2014; Kazdin, 1980; Kratochwill et al., 2010). As a consequence, researchers are encouraged to reflect carefully on the practical constraints of the context in which the study is conducted, on the desired design features (e.g., minimum phase lengths), and then tailor the randomization strategy to meet these constraints. Restricted randomization schemes have been developed to ensure that: (1) the desired number of phases and minimum phase lengths are included in reversal designs (Onghena, 1992); (2) the treatment alternates quickly enough in an alternating treatment design (Onghena & Edgington, 1994); (3) the baseline series stabilizes prior to commencement of the intervention phase (Ferron & Ware, 1994); (4) the intervention start points are staggered by a minimum amount of time in multiple-baseline designs (Koehler & Levin, 1998), and (5) researchers are able to obtain visually acceptable patterns by extending phases in multiple-baseline designs (Ferron & Jones, 2006) and reversal designs (Ferron & Levin, 2014).

The present Monte Carlo simulation study employs nonparametric randomization tests in the company of a recently proposed methodological addition that greatly enhances the internal validity of AB and ABAB single-case intervention designs (Ferron & Levin, 2014; Levin, Evmenova, & Gafurov, 2014). In these designs, A typically represents a baseline, control, or standard treatment phase containing repeated outcome measurements and B represents an intervention, experimental, or new treatment phase also containing repeated outcome measurements. Here we examine the methodological addition's effect on the statistical conclusion validity (manifested by both Type I error control and increased statistical power) of randomization tests in single-case AB and ABAB designs, in both their single-case ($N = 1$) and multiple-case ($N > 1$) forms. In the following section, we first describe the methodological addition that enhances the internal validity (scientific credibility) of single-case intervention research and then outline how the addition is incorporated into a randomization test to improve the test's statistical conclusion validity. Our decision to start our investigations with a single-participant (N = 1) AB design was not because we are advocating for the use of such a design, but because it provides the simplest point to begin study of the impact of the methodological addition. Once we have established the effects on statistical conclusion validity in the simplest situation, we will progressively add complexities to strengthen the design, building to the multiple-participant (N > 1) ABAB design.

Edgington's (1975) Random Intervention Start-Point Model

Of four different types of randomization that can be incorporated into randomization in single-case AB experimental studies (specifically, within-case phase randomization, between-case intervention randomization, case randomization, and intervention start-point randomization (see Ferron & Levin, 2014), the last, highly creative, type was originally developed by Edgington (1975) and requires that the researcher: (1) randomly select an intervention start point from two or more that had been previously deemed acceptable; and then (2) assign to the case the start point that was actually selected. Although not applied in the conventional treatment randomization manner, this unique form of randomization increases a single-case study's internal validity and, when accompanied by the statistical test described in the following paragraph, it can increase the study's statistical conclusion validity as well. Moreover, this randomized intervention start-point approach can function to provide a true (i.e., scientifically credible) experimental comparison of two or more intervention (or intervention and control) conditions based on either one case or multiple cases per condition (for examples and discussion, see Ferron & Levin, 2014; Koehler & Levin, 1998; Levin, Lall, & Kratochwill, 2011; Levin & Wampold, 1999; and Marascuilo & Busk, 1988).

With the randomized intervention start-point model, a randomization statistical test is conducted on the difference between the means of all B and all A series outcomes for each of the intervention start-point divisions (or *transitions*) that could have resulted from the random-selection process (see also Edgington & Onghena, 2007). [Moreover, any other summary measure of relevance to the researcher's hypothesis about the nature of change from Phase A to Phase B (e.g., change in the series' medians, slopes, variances) can also be the focus of a randomization-test analysis.]

With the resulting set of mean differences yielding a randomization distribution, the mean difference associated with the actual intervention start point is examined to see where it falls within the set. The probability of obtaining a mean difference as extreme as or more extreme than the actual mean difference represents the unlikelihood of the outcome. Either signed or unsigned mean differences are considered for one- and two-tailed hypothesis tests, respectively. For example, for an AB design with one case, 25 outcome-assessment periods, and 20 potential intervention start points, if the actual start point were found to produce the largest mean difference (in the predicted direction) between the B and A series outcomes, then the one-tailed significance probability associated with that event would be given by $p = 1/20 = .05$. For a two-tailed test, as or more extreme opposite-sign mean differences would also need to be taken into account. For instance, if there were a mean difference equal in magnitude but opposite in sign to the one just indicated for the actual intervention start point, then the two-tailed significance probability would be 2/20 = .10.

In Edgington's (1975) random intervention start-point model for a one-case AB design, it is assumed that the A phase consists of a baseline series, the B phase consists of an intervention series, and that the former logically precedes the latter. With those assumptions, the number of possible outcomes (B−A mean differences) in the randomization distribution is k, the number of potential intervention start points. Accordingly, with one case, 30 total observations, and $k = 10$ potential intervention start points, if the actual B−A mean difference produced were the largest of the 10 and in the predicted direction, then the one-tailed significance probability of that outcome would be $p = 1/10 = .10$. In order to achieve statistical significance at a traditional $\alpha = .05$ level (one-tailed), one would need to include at least $k = 20$ potential intervention start points in the randomization distribution (i.e., so that if the most extreme mean difference in the predicted direction were obtained, then p would equal 1/20 = .05). To achieve statistical significance with $\alpha = .05$ via a two-tailed test, a longer series with a minimum of k = 40 potential intervention start points would be required (i.e., so that $p = 2/40 = .05$ is possible).

Randomized Order (Dual Randomization) Addition to the Edgington Model

Edgington (1975) proposed his random intervention start-point design-and-analysis procedure 40 years ago. It has been incorporated into a variety of single-case intervention designs (e.g., Koehler & Levin, 1998; Levin & Wampold, 1999; Marascuilo & Busk, 1988; Onghena, 1992) and is being implemented in its original form to this day. However, it will be shown here that an addition to the procedure (referred to here as a *modified* procedure), which enhances its internal validity by eliminating bias due AB phase-order effects, is possible and one that is applicable in a number of single-case intervention investigations. To illustrate, suppose that instead of A representing a baseline or control phase, it represents one type of experimental intervention—say, a behavioral intervention for combatting a particular phobia. In contrast, B might represent a cognitive intervention targeting the same phobia. Within that context, the case receives both interventions. To have a legitimate (unconfounded) comparison of Intervention A and Intervention B, it is imperative that the order in which the two interventions are administered to the case is randomly (rather than arbitrarily) determined. The preceding statement applies whether the investigation includes only one case or multiple cases (although in multiple-case situations, systematic counterbalancing of intervention orders across cases might be implemented to achieve the same goal).

In addition, it is worth noting that A and B need not refer only to two competing interventions. Rather, suppose that A represents a baseline, standard, or control condition and B an intervention condition. As has been suggested previously (e.g., Kratochwill & Levin, 2010), further suppose that prior to the commencement of the actual experiment, a mandatory baseline (or adaptation/warm-up) phase (A') is required of all cases. With A' included, it would then be possible, appropriate, and presumably acceptable to researchers to begin the experiment proper by randomizing each case's subsequent A and B phases (i.e., an A randomly selected to be first means that the case remains in the baseline condition, followed by the B intervention condition; and a B randomly selected to be first means that the case begins with the intervention condition, followed by the A baseline condition). Accordingly, the modified order-randomization procedure is applicable in either one- or two-intervention AB designs, with the prospect of improving both design (internal validity) and analysis (statistical-conclusion validity) of two-phase single-case intervention studies.

With intervention-order randomization built into the just-discussed one-case example based on 30 total observations and 10 potential intervention start points, in addition to the intervention start points associated with the conventional AB order of intervention administration, one would also need to consider the possibility that Intervention B had been randomly selected to be administered first. If that had happened, there would be a corresponding 10 potential intervention start points for the BA order of intervention administration, resulting in a total of $k = 20$ potential start-point outcomes that would be included in the complete randomization distribution.

Multiple-Case Extension of the Modified Edgington Model

As we will show, the order-randomization procedure applies to multiple-case (*replicated*) AB situations as well, increasing the total number of possible randomization-distribution outcomes by a factor of 2^N, where N represents the number of cases. Specifically, with N cases and one of k_i potential intervention start points randomly selected for each case, with Marascuilo and Busk's (1988) multiple-case extension of Edgington's (1975) single fixed-order intervention start-point model, a total of $\prod_{i=1}^{N} k_i$ randomization-distribution outcomes are possible, and in the special case for which all k_i are equal to k, this quantity reduces to k^N. With the addition of an order-randomization process to create the present dual randomization model, the total number of possible randomization-distribution outcomes increases to $\prod_{i=1}^{N} k_i \times 2^N$ and $k^N \times 2^N = (2k)^N$ for the general and special-case situations, respectively.

Hypothetical example We illustrate the present random-order randomization-test procedure for a replicated single-case AB design by means of a hypothetical example. Suppose that a language researcher wishes to improve the baseline vocalization output (A phase) of two low word-producing children through some type of positive-reinforcement intervention (B phase). For the random-order version of the present example we assume that a mandatory A' baseline (warm-up) phase was initially administered, followed by a random determination of whether the first phase of the actual study would be a baseline (A) or an intervention (B) phase, thereby producing either an A'AB or A'BA design. Although in comparison to a traditional fixed-order AB design, this type of randomized AB design is more scientifically credible (especially when replicated across cases), the latter design was not considered in the current What Works Clearinghouse (WWC) single-case intervention design *Standards* (Kratochwill et al., 2010). Our hypothetical study is presented simply to illustrate both the original (Edgington, 1975) fixed-order and the present random-order randomization-test procedures, without taking into account the study's internal-validity characteristics. Consideration of internal-validity issues is included later in the Discussion section.

In this hypothetical study, the number of single-word vocalizations by each child during a 5-minute play period is recorded, with Child 1 observed in each of 25 daily sessions and Child 2 observed in each of 15 daily sessions, and where both children must be observed in at least 3 A sessions and 3 B sessions (thereby resulting in 20 and 10 potential intervention transition points for Child 1 and

Child 2, respectively). In addition, because the researcher wishes to randomize the intervention order (AB or BA) for each child, three preliminary five-minute A' warm-up sessions are provided prior to the start of the children's actual experimental sessions. An initial coin toss determines that Child 1 will be administered an AB intervention order, with the 20 potential intervention transition points specified from between the 4th and 23rd sessions inclusive and the randomly selected actual intervention transition point occurring just prior to Session 10. For Child 2, a BA intervention order results from a second coin flip, with the 10 potential intervention transitions specified from between the 4th and 13th sessions inclusive and an actual randomly selected intervention transition point just prior to the 7th observation.

The A- and B-phase observations are presented in Table 1. Given the present random-order AB intervention start-point randomization model, the data were analyzed with Gafurov and Levin's (2014) single-case *ExPRT (Excel® Package of Randomization Tests*) package—see Levin et al. (2014) for complete information about *ExPRT*. In Table 2 are presented the B−A mean differences associated with each of the potential intervention transition points for the two children.

The first Table 2 entry of 2.41 for Child 1, which corresponds to an A-to-B intervention transition point just prior to Observation 4, was calculated by taking the average of Child 1's Observations 4 through 25 (mean B phase = 6.41) minus the average of that child's Observations 1 through 3 (mean A phase = 4.00). The same process was followed for each of the subsequent 19 potential intervention points for Child 1, which ends with the average of that child's Observations 23 through 25 (mean B phase = 8.00) minus the average of that child's Observations 1 through 22 (mean A phase = 5.86), resulting in Child 1's final mean difference of 2.14 in Table 2. Next, and as indicated in Table 2's Footnote a, 20 additional mean differences were calculated for Child 1 under the assumption that instead of an A−B intervention order, the reverse B−A order had been selected. Under that assumption, the first mean difference for Child 1 would be 4.00 − 6.41 = −2.41, which is exactly the same numerically but opposite in sign to the previously calculated child's first value in Table 2. The same is true for all of Child 1's calculated reverse-order values, including the 20th one, which is now −2.14. The same process applied to Child 2's data yields the 10 actual B−A mean differences presented in Table 2 (i.e., 6.00 − 4.92 = 1.08 for the first one), as well as 10 reverse-order and opposite-sign A−B mean differences.

Table 1. Hypothetical data for Child 1's 25-observation series, with a randomly selected AB intervention order, 20 potential intervention transition points (between Observations 4 and 23 Inclusive), and a randomly selected actual intervention transition point just prior to Observation 10; and for Child 2's 15-observation series, with a randomly selected BA intervention order, 10 potential intervention transition points (between Observations 4 and 13 Inclusive), and a randomly selected actual intervention transition point just prior to Observation 7

	Child 1			**Child 2**	
Observation	**Phase**	**Vocalizations**	**Observation**	**Phase**	**Vocalizations**
1	A	4	1	B	6
2	A	3	2	B	5
3	A	5	3	B	7
4	A	5	4	B	5
5	A	2	5	B	6
6	A	5	6	B	5
7	A	3	7*	A	4
8	A	4	8	A	5
9	A	4	9	A	3
10*	B	5	10	A	5
11	B	6	11	A	4
12	B	7	12	A	5
13	B	6	13	A	6
14	B	7	14	A	5
15	B	8	15	A	6
16	B	7			
17	B	9			
18	B	8			
19	B	6			
20	B	8			
21	B	9			
22	B	8			
23	B	7			
24	B	9			
25	B	8			

*Actual intervention transition point.

Table 2. The B−A mean difference associated with: (1) each of Child 1's 20 potential intervention transition points (O_4-O_{23}) for a randomly selected AB intervention order; and (2) each of Child 2's 10 potential intervention transition points (O_4-O_{13}) for a randomly selected BA intervention order

	Child 1	**Child 2**
Potential Intervention Point	**B-A Mean Difference**[a]	**B-A Mean Difference**[b]
O_4	2.41	1.08
O_5	2.23	0.84
O_6	2.90	1.00
O_7	2.79	0.89*
O_8	3.14	0.55
O_9	3.30	0.52
O_{10}	3.49*	-0.06
O_{11}	3.53	-0.10
O_{12}	3.46	-0.50
O_{13}	3.28	-0.67
O_{14}	3.29	
O_{15}	3.19	
O_{16}	2.97	
O_{17}	2.94	
O_{18}	2.58	
O_{19}	2.41	
O_{20}	2.69	
O_{21}	2.60	
O_{22}	2.24	
O_{23}	2.14	

*Mean difference associated with the actual intervention transition point. [a] The 20 A−B mean differences are also calculated and added to these to form a 40-outcome randomization distribution; all of the A−B mean differences are the same as the corresponding B-A mean differences given here but opposite in sign. [b] The 10 A−B mean differences are also calculated and added to these to form a 20-outcome randomization distribution; all of the mean A−B differences are the same as the corresponding mean B−A differences given here but opposite in sign.

The resulting joint randomization distribution therefore contains 40 mean differences for Child 1 combined with 20 mean differences for Child 2, for a total of 40 × 20 = 800 averaged mean differences (i.e., Child 1's 1st mean difference averaged with Child 2's 1st mean difference, Child 1's 1st mean difference averaged with Child 2's 2nd mean difference, all the way up to and including Child 1's 40th mean difference averaged with Child 2's 20th mean difference). When that is done by the *ExPRT* program, it is found that the *actual* joint mean difference that was obtained in the study is 2.19, which is Child 1’s mean difference associated with that child’s actual intervention transition point of O_{10} (3.49) averaged with Child 2’s actual intervention transition-point mean difference of O_7 (.89). Of the 800 outcomes in the joint randomization distribution, a value of 2.19 is the 10th highest, which results in a one-tailed significance probability of $p = 10/800 = .0125$. For this example, had a one-tailed Type I error probability (α) of .05 been selected, it could be concluded that the positive-reinforcement intervention (B) distribution values differed statistically from those in the baseline distribution (A), with the additional inference that the former distribution’s values were higher. We note that both here and in the various simulations conducted in the present series of investigations, one-tailed tests are conducted because it is assumed that [especially in single-case A (baseline) − B (intervention) research] the researcher has a clear and defensible rationale for the direction of change that is associated with the intervention.

Insofar as randomization tests are not tailored to test for the equality of two populations’ specific parameters, all that can be tested for is the equality of the two population distributions *per se*. For the present randomization test, the test statistic involves sample-mean differences and because that is the test that produced a statistically significant result here (favoring the intervention phase over the baseline phase), a reasonable inference is that there was an A- to B-phase upward shift in the children’s level of responding.

Advantages of the Order Randomization Modification

The present order-randomization approach enhances the internal validity of a single-case AB design by virtue of its removing bias stemming from intervention-order effects. As an important byproduct, the approach also elevates the status of the basic AB single-case intervention design from a WWC *Standards* “acceptable design” standpoint (Kratochwill et al., 2010), particularly when replicated across independent participants at different points in time. According to the WWC *Standards*, two-phase A (Baseline) – B (Intervention) designs are not scientifically credible (and therefore unacceptable) because they suffer from too many potential sources of internal invalidity. For extended discussion of acceptable designs, see Kratochwill, et al. (2010, 2013).

Including outcomes from both intervention-administration orders in the randomization distribution also provides fundamental pragmatic advantages for single-case intervention researchers. First, with the original Edgington (1975) model, a researcher would need to designate 20 potential intervention start points (based on at least 21 total observations) to produce a randomization test that is capable of detecting an intervention effect with a one-tailed Type I error probability less than or equal to .05. With the present procedure, a researcher would need to designate only half as many potential intervention start points (here, 10, based on a total of 11 total observations, resulting in 20 possible outcomes) to detect an intervention effect. A related reason why the present procedure has practical importance for single-case intervention researchers is that (and as will be demonstrated here) relative to the original Edgington (1975) model, the modified approach may produce statistical-power advantages as well. Thus, for no more expense than a coin to flip, a researcher might reap both methodological and statistical benefits by adopting the present dual-randomization procedure rather than either the original single-randomization Edgington model or Marascuilo and Busk's (1988) multiple-case extension of it.

Relationship to Traditional Experimental Designs and Statistical Analyses

Although unrecognized at the time that the present order-randomization approach was initially conceptualized, its logic maps directly onto a statistical procedure in the traditional group randomized treatment-design literature. In particular, consider a randomized two-treatment correlated-samples (or within-subjects) design based on *N* participants, to which a nonparametric randomization test is applied as an appropriate alternative in (especially small-sample) situations where the normality assumption of a correlated-samples *t* test (or a one-sample repeated-measures analysis) is questionable.

To illustrate that situation, we revisit an example that was recently presented by Ferron and Levin (2014, p. 174). Suppose that in a sample of $N = 8$ adults, each participant is administered two different fear-reducing treatments, A (a behavioral treatment) and B (a cognitive intervention), with the former posited to be more effective than the latter. It is determined in advance that the equal-effectiveness hypothesis will be tested with a randomization test based on a one-tailed α of .05. To produce a scientifically credible experiment, the order in which the two treatments are administered is again randomly determined on a case-by-case basis by means of coin flips: say, heads represents an AB order and tails a BA order. On the basis of that process, let us suppose that 5 participants ended up in the AB condition and 3 in the BA condition. Following the administration of each treatment, participants' fear responses are assessed on a 7-point Likert scale, with higher numbers indicating greater fear. With the measure of interest defined as the difference between each participant's B and A ratings (i.e., B−A), the following outcomes were obtained for the 8 participants:

+3.0 +3.5 −1.5 +2.0 +4.5 +3.5 −2.0 +4.0

The observed test statistic is given by the average of these differences, which is equal to +17/8 = 2.125. A randomization distribution is created from the $2^N = 2^8 = 256$ possible ways in each + and − signs could be attached to these 8 numerical values. For example, the first outcome in the randomization distribution (with all + signs) would be:

+3.0 +3.5 +1.5 +2.0 +4.5 +3.5 +2.0 +4.0

yielding a mean difference of +24/8 = 3.000, and the last (with all minus signs) would be:

−3.0 −3.5 −1.5 −2.0 −4.5 −3.5 −2.0 −4.0

yielding a mean difference of −24/8 = −3.000. The remaining 254 possible outcomes would fall somewhere between these two extremes.

The actually obtained mean difference of +2.125 appears to be on the higher side of this distribution. In fact, it turns out to be among the 9 highest possible outcomes (specifically, an outcome that is exceeded by only 5 outcomes and that is tied with 3 others). Accordingly, a one-tailed test of the hypothesis that the A and B treatments have equal distributions would be associated with a p-value (consistent with the alternative hypothesis that Treatment B is producing higher fear ratings than Treatment A) that is equal to 9/256 = .035. Because this value is less than the predetermined α of .05, it would be concluded that the actually obtained mean difference of +2.125 is statistically significant.

Note that for this conventional-group design and associated randomization test, the all-possible assignment of + and – signs to the 8 absolute B−A differences corresponds exactly to the logic and operationalization of the single-case AB order-randomization procedure to be investigated here. In particular, the procedure incorporates two separate forms of randomization for each of the N participating cases, Edgington's intervention start-point randomization and AB order randomization. In the simplest situation where there is only one potential intervention start point for each case (as in the just-presented $N = 8$ example), the total number of possible start-point randomizations is equal to $k^N = 1^8 = 1$. The present order-randomization procedure involves each of the 8 participants contributing two differences (i.e., B−A and A−B) to the randomization distribution, resulting in $2^N = 2^8 = 256$ joint randomization outcomes, and which, according to the previously given special-case dual-randomization formula, $k^N \times 2^N$, yields a total of $1 \times 2^8 = 256$ possible randomization outcomes. This total is identical to the number of possible randomization-distribution outcomes associated with the just-presented example. It is instructive to note that the total number of possible randomization outcomes associated with order randomization can be alternatively expressed as $\sum_{x=0}^{N} \binom{N}{x}$, where N = the number of cases and x = the number of positive B−A differences that could be associated with the N actual outcomes. For the present example, this expression is equal to $\sum_{x=0}^{8} \binom{8}{x}$, or

$$\binom{8}{0} + \binom{8}{1} + \binom{8}{2} + \binom{8}{3} + \binom{8}{4} + \binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8}$$
$$= 1 + 8 + 28 + 56 + 70 + 56 + 28 + 8 + 1$$
$$= 256$$

Thus, when there is only one potential intervention point for each case and the AB design includes multiple observations, the present randomized-order test based on the difference between the A- and B-phase means maintains the same correspondence with a conventional-group correlated-samples randomization test as was shown here. Implicit in the conventional correlated-samples test is that with random assignment to treatment conditions, outcomes representing both orders of treatment administration need to be considered in the randomization test distribution. As such, the present order-randomization procedure is not really a special case at all, but rather the single-case analog of a correlated-samples randomization t test.

Focus of the Present Investigations

The focus of our series of simulation investigations was to examine the Type I error and statistical power characteristics of the dual-randomization modification (intervention start-point *plus* intervention order) relative to those of Edgington's (1975) and Marascuilo and Busk's (1988) original single-randomization (intervention start-point) test procedures. In this study we present randomized intervention-order findings not just for a basic two-phase AB design, but also for a randomized pairs variation of that design (Levin & Wampold, 1999), a single-case adaptation of the conventional-group crossover design, and Onghena's (1992) four-phase ABAB design.

Investigations 1-3: Randomized Intervention Order for the Basic AB Design

Investigation 1

Method In Investigation 1, the focus was on 30-observation designs for a single participant (i.e., $N = 1$), where the intervention start point was randomly selected from the middle 20 observations. The series length of 30 was chosen for initial examination because: (1) 20 start points is the minimum number needed to obtain a statistically significant result with a one-tailed α of .05 for an AB randomized start-point design with one case; and (2) the WWC *Standards* require a minimum of five observations in each phase (Kratochwill et al., 2010, 2013).

Data were generated using SAS IML (SAS, 2013), where the time-series data were obtained by adding an error vector to an effect vector. The error vector was created such that it was distributed normally and had an autocorrelation of 0 or .3 by using SAS's autoregressive moving-average simulation function (ARMASIM). The autocorrelation values of 0 and .3 were motivated by a survey of actual single-case studies where it was reported that the average autocorrelation was .2, after adjusting for bias in the estimates (Shadish & Sullivan, 2011). To obtain simulated errors based on an autocorrelation of .3, the autoregressive parameter matrix was set to {1 −.3}, the moving average parameter matrix was set to {1 0}, and a standard deviation of the independent portion of the error was set to 1.0 (for details on the simulation algorithm see Woodfield, 1988). The effect vector was coded to have values of 0 for all baseline observations, and values of *d* for all intervention phase observations, and thus *d* corresponds to the mean shift between intervention and baseline observations in standard deviation units, $(\mu_B - \mu_A)/\sigma$ (see Busk & Serlin, 2005), where the standard deviation is based on the independent portion of the within-case error term (see, for example, Levin, Ferron, & Kratochwill, 2012) (for an alternative operationalization of d that corresponds mathematically to a conventional groups effect-size measure, see Shadish et al. (2014)). The value of *d* was varied to examine the one-tailed Type I error probability for $d = 0$ and the powers for *d*s ranging from .5 to 5 in increments of .5. For reference, if the *d* used for the present data generation is estimated for each of the 200 Phase A-to-Phase B contrasts examined in the survey of single-case interventions reported by Parker and Vannest (2009), the empirically observed values of *d* (assuming no autocorrelation for simplicity) for the 10th, 50th, and 90th percentile ranks are estimated to be 0.46, 1.70, and 3.88, respectively.

By crossing each design (single, dual), with each level of autocorrelation ($r = 0$, .3), and each effect size ($d = 0$ to 5, in increments of .5), 2 × 2 × 11 = 44 conditions were obtained, and for each of these conditions the data for 10,000 studies were simulated. The data for each simulated data set were analyzed using a randomization test in which the obtained test statistic ($M_B - M_A$) was compared to the complete randomization distribution. The proportion of simulated studies in which the randomization test led to a one-tailed *p*-value of .05 or less was determined to estimate the rejection rate (Type I error or power) of the randomization test for each of the 44 experimental conditions.

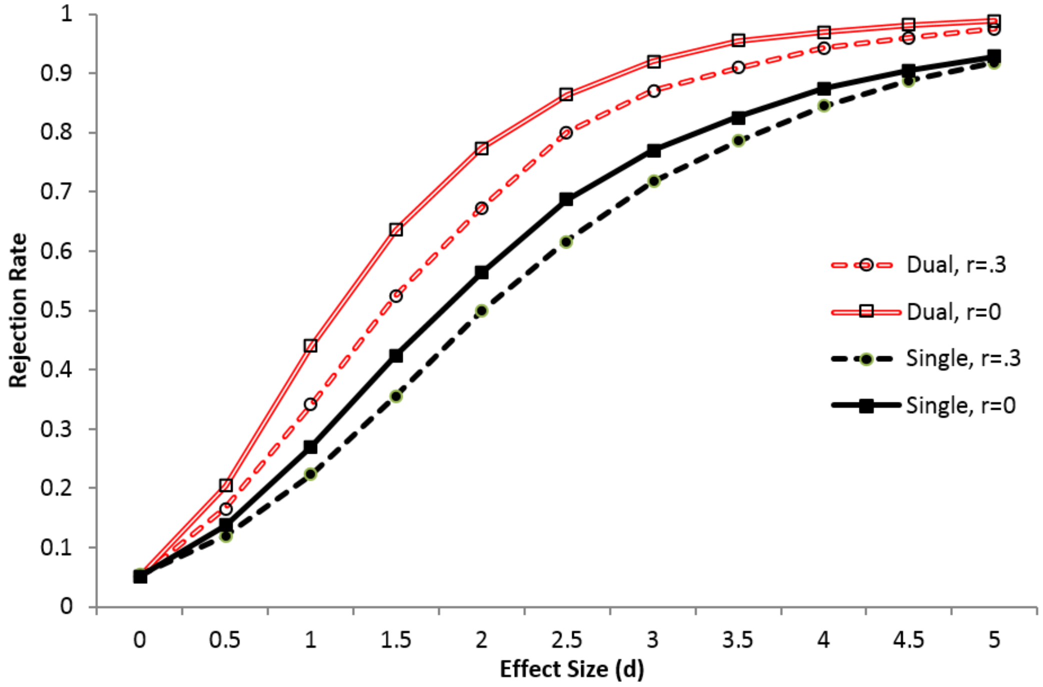

Figure 1. Investigation 1: Comparison (α = .05, one-tailed) of randomization tests for a one-case (*N* = 1) AB randomized intervention start-point design (Single) and the randomized intervention start-point plus randomized intervention- order design (Dual), where the start point was randomly selected between the 6th through the 25th observations inclusive in a 30-observations study. The rejection rate of the null hypothesis is shown as a function of the effect size and level of autocorrelation.

Results Results are shown in Figure 1 for Edgington's (1975) original procedure (single) and for the present randomized-order modification (dual). As may be seen in that figure, when the effect size is 0, all situations are associated with empirical powers (which, for $d = 0$ are equivalent to Type I error probabilities) that correspond to their nominal .05 values. Not surprisingly, based on previous findings (e.g., Ferron & Sentovich, 2002; Ferron & Ware, 1995; Levin et al., 2011), it may also be seen that for *ds* > 0 power is uniformly higher for $r = 0$ than for $r = .3$. As the effect size increases, so does power, although more rapidly for the dual-randomization procedure than for its single-randomization counterpart. The largest power differences, favoring the former, reach .21 in the $r = 0$ situation for *ds* of 1.5 and 2.0; and in the $r = .3$ situation the largest power difference is .18 for a *d* of 2.5.

Investigation 2

Method In Investigation 2, series length (i.e., the number of observations) was systematically varied for a single-participant ($N = 1$) design, while holding the effect size constant at $d = 2$. A *d* of 2 was chosen because it is a large enough effect to typically be of interest to a single-case researcher. Yet, a *d* of 2 is small enough that it is not readily detectable (power < .80) in a single-participant 30-observations design when there is a moderate autocorrelation of .30 and applying either the single- or dual-randomization approach (as may be seen in Figure 1, where powers are .50 and .67, respectively). The simulation methods paralleled those of the initial investigation (including a one-tailed α of .05), but *d* was held constant at 2.0 for all conditions and series length was varied from 20 to 150 in increments of 10. The number of potential intervention start points was always the series length minus 10 to ensure at least five observations in the baseline and intervention phases.

Results Results for this set of simulations are provided in Figure 2, where with an autocorrelation of .30, power of at least .80 is attained for the dual-randomization approach with 60 observations (power = .81), in contrast to the single-randomization design where .80 power is not quite attained even with 150 observations (power = .79). For 30 to 100 observations, the power difference between the two randomization schemes (favoring dual) ranges from .13 to .31 when the autocorrelation is 0 and from .17 to .30 when the autocorrelation is .30.

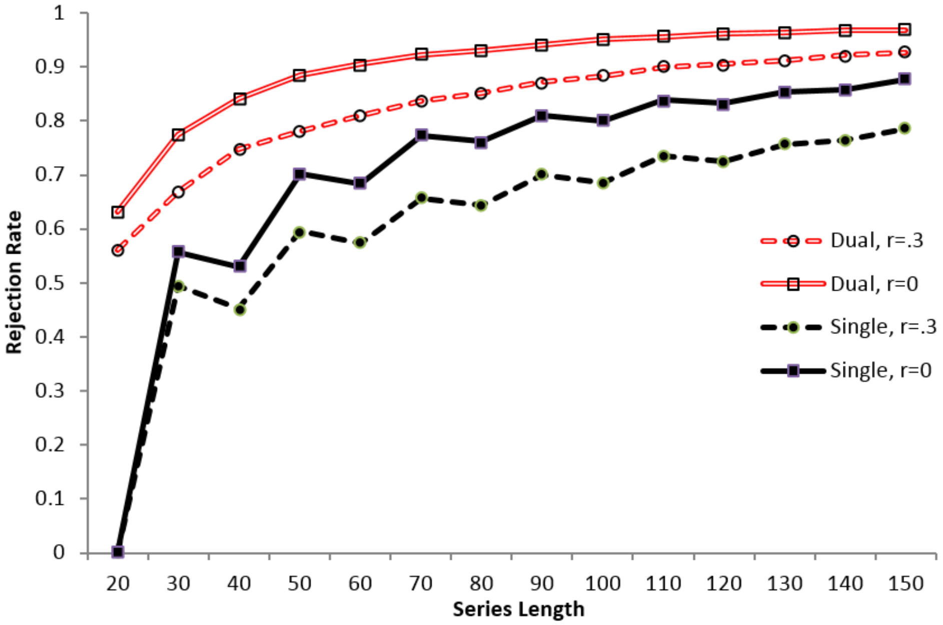

Figure 2. Investigation 2: Comparison (α = .05, one-tailed) of randomization tests for a one-case (*N* = 1) AB randomized intervention start-point design (Single) and the randomized intervention start-point plus randomized intervention-order design (Dual). The rejection rate of the null hypothesis is shown as a function of series length and level of autocorrelation. The effect size is 2.0 and the number of potential intervention start points (*x*) is equal to the series length minus 10 and encompasses the middle *x* observations.

It should be noted that the power is 0 for the single-randomization scheme with 20 observations because there are only 10 possible intervention start points and thus statistical significance cannot be obtained at the one-tailed .05 level. In addition, the undulation in the power curves for the single-randomization approach makes sense when one recognizes that: (1) for a series length of 30, statistical significance with α = .05 can be attained only for the most extreme of the 20 permutations; and (2) with a series length of 40, statistical significance can again be attained only for the most extreme permutation, but now there are 30 permutations and so the most extreme is somewhat more difficult to achieve. Although power drops for the 40-observation series, with a series length of 50, statistical significance can be attained for either of the two most extreme permutations and thus power jumps back up again.

Investigation 3a

Method In Investigation 3a, the effect of multiple-case replications (i.e., $N > 1$) on the power of the single- and dual-randomization procedures was examined. More specifically, a design with 15 observations and 5 potential intervention start points, randomly selected from observations 6 through 10, was examined with 2, 3, 4, 5, and 6 participants based on a one-tailed α of .05. For the single-randomization approach, 7 and 8 participants were also included. These numbers of participants seemed reasonable given the survey by Shadish and Sullivan (2011), in which it was found that the number of cases in single-case studies averaged 3.64, with a range of 1 to 13. In the present study, effect sizes varied from 0 to 3 in increments of .5 and the autocorrelation was set either to 0 or .3.

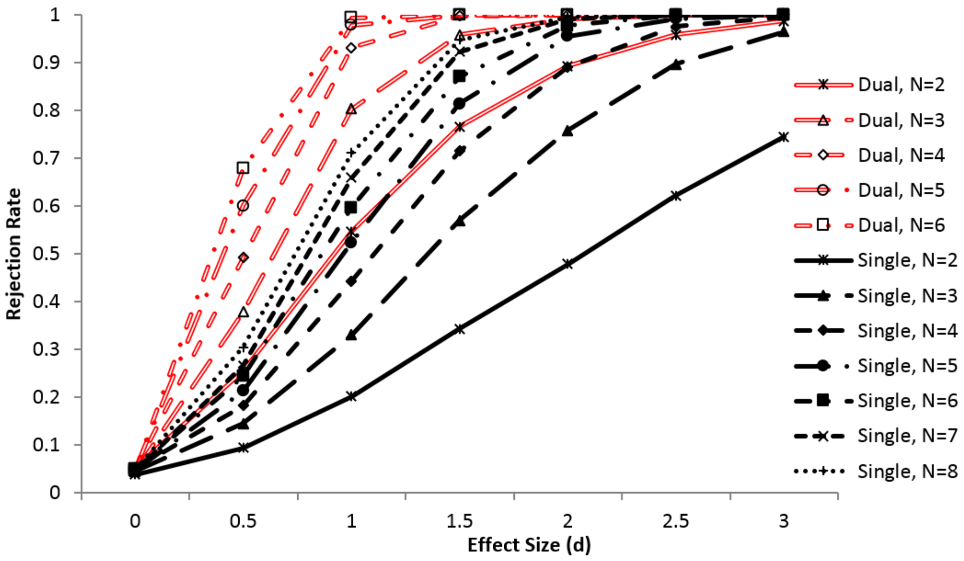

Figure 3. Investigation 3a: Comparison (α = .05, one-tailed) of randomization tests for the Single and Dual basic AB randomized designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15-observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of 0.

Results Results from simulations where the autocorrelation is 0 are shown in Figure 3, whereas those for an autocorrelation of .3 are shown in Figure 4. In both figures, it may be seen that for all sample sizes the empirical Type I error probabilities are well controlled at .05 for both the single- and dual-randomization approaches. The important thing to note is that in both figures, for all effect sizes the dual approach based on as few as *N* = 3 participants has associated power that is greater than or equivalent to the single approach based on *N* = 8 participants. For example, in Figure 4 it may be seen that with an autocorrelation of .3, *N* = 3 dual- and *N* = 8 single-randomization powers are .66 and .61, respectively, for an effect size of 1.0; and they are .90 and .89, respectively, for an effect size of 1.5.

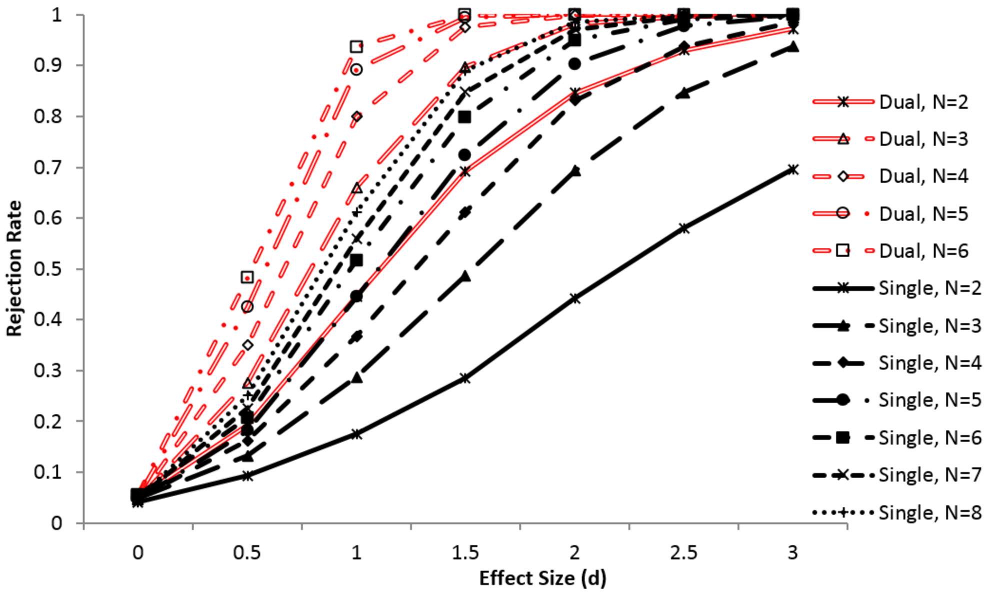

Figure 4. Investigation 3a: Comparison (α = .05, one-tailed) of randomization tests for the Single and Dual basic AB randomized designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of .3.

Investigation 3b

Method In this investigation, the simulations of Investigation 3a were replicated with the sole difference being that a two-tailed test with α = .05 was conducted, as opposed to a one-tailed test.

Results The results are summarized in Figure 5 for an autocorrelation of 0 and in Figure 6 for an autocorrelation of .3. Again, it may be seen that all of the empirical Type I errors are at the expected .05 level for both autocorrelation values. Although the Investigation 3a results (i.e., the equivalence of dual-randomization *N* = 3 and single-randomization *N* = 8) were not identical here, the general pattern was. In this case, however, the appropriate power equivalence involves dual *N* = 4 and single *N* = 8. Specifically, in Figure 6 it may be seen that with an autocorrelation of .3, the former and latter powers are .65 and .61, respectively, for an effect size of 1.0; and they are .93 and .89, respectively, for an effect size of 1.5.

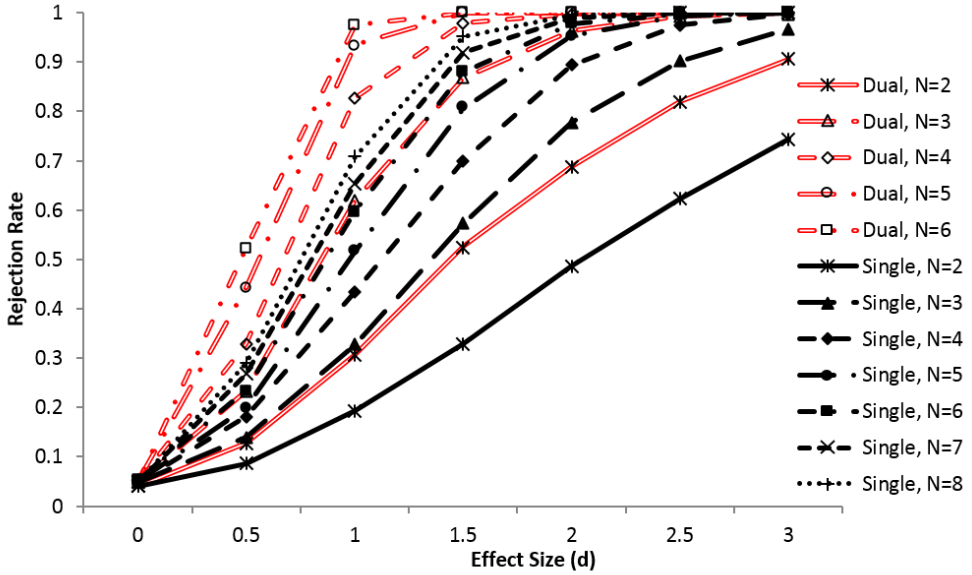

Figure 5. Investigation 3b: Comparison (α = .05, two-tailed) of randomization tests for the Single and Dual basic AB randomized designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of 0.

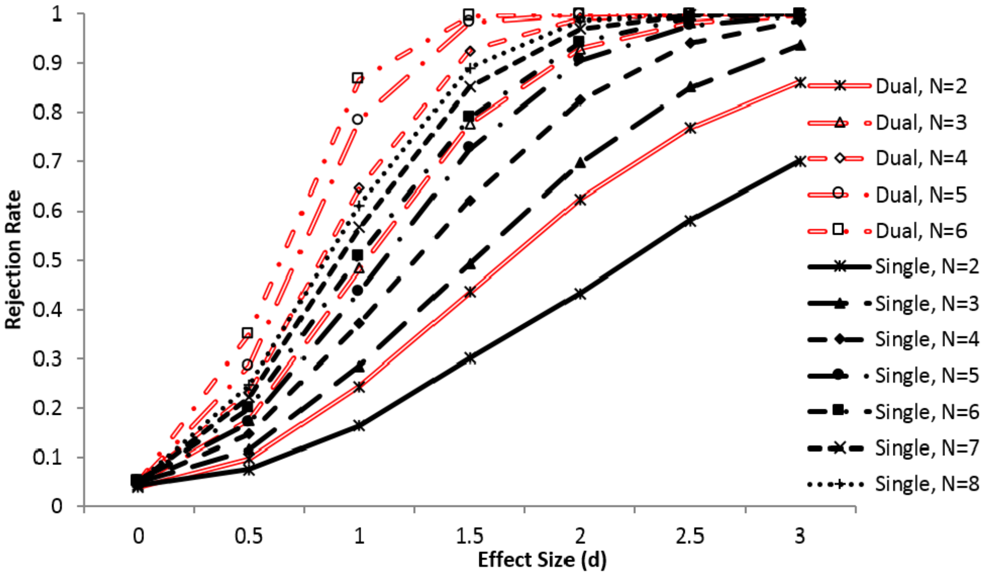

Figure 6. Investigation 3b: Comparison (α = .05, two-tailed) of randomization tests for the Single and Dual basic AB randomized designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of .3.

Thus, in the present investigation we observe that for two-tailed tests the dual-randomization power benefits (relative to single randomization) are comparable to those reported for Investigation 3a's one-tailed tests. It is important to point out, however, that the situations examined here were all based on multiple-case ($N > 1$) designs. It turns out that for the special-case $N = 1$ situation, although the dual- over single-randomization power advantage is evident when one-tailed tests are conducted (as was true in Investigations 1 and 2), the dual- and single-randomization schemes yield equivalent power results with two-tailed tests. Because the two-tailed test is based on randomization-distribution absolute-value outcomes, the dual-randomization distribution contains every outcome of the single-randomization distribution as well as its opposite-order complementary outcome, thereby yielding exactly the same p-value for each test. (To illustrate these notions, see Child 1's hypothetical data, including Footnote a in Table 2. The 40 unsigned mean differences (i.e., 20 |B−A| plus 20 |A−B|) would constitute the dual-randomization distribution for a two-tailed test). Because there are across-case combinations when $N > 1$, there is no longer a one-to-one correspondence between the single- and dual-randomization distributions and so their powers will generally differ, with the latter being greater (as was observed in Figures 5 and 6).

Investigation 4: Randomized Intervention Order and/or Randomized Intervention Assignment in Levin and Wampold's (1999) AB Pairs Design

Another type of dual-randomization strategy is possible when a case consists of a pair of participants, as in Levin and Wampold's (1999) simultaneous intervention start-point model. With the Levin-Wampold model, *N* participant (or other unit) pairs are created and the members of each pair are randomly assigned to two different intervention conditions (or to an intervention and control condition), X and Y. With this model, Levin and Wampold presented two hypotheses that would be of interest to researchers: (1) a *general intervention effectiveness* hypothesis, namely that averaged across the two intervention conditions, there is no difference between Phase A and Phase B performance (analogous to the time main effect in a conventional two-treatment pretest-posttest design); and (2) a *comparative intervention effectiveness* hypothesis, namely that the change in participants' performance from Phase A to Phase B is the same in the two intervention conditions (analogous to the treatment-by-time interaction in a conventional two-treatment pretest-posttest design). Unrecognized by Levin and Wampold at the time, the randomization test of each of these hypotheses could potentially benefit from an additional randomization component. For the general intervention effectiveness hypothesis, that component is AB order randomization of the kind that we have considered in Investigations 1-3, either with or without a mandatory A' baseline phase; and for the comparative intervention hypothesis, that component consists of within-pair intervention randomization, wherein pair members are randomly assigned to the two intervention conditions.

Implementing either of these randomization types increases the total number of possible outcomes from $\prod_{i=1}^{N} k_i$ for Levin and Wampold's (1999) original single randomization-test procedure (i.e., the number of potential intervention start points for each pair) to $2^N \times \prod_{i=1}^{N} k_i$ for the present dual approach (i.e., either the number of possible random assignments of AB orders or the number of possible random assignments of interventions to pair members, times the number of potential intervention start points for each pair). In Investigation 4, we examine the statistical power consequences associated with the dual approach's additional randomization component, for both the general and the comparative intervention effectiveness hypotheses.

Method

A power comparison of dual versus single randomization for the two hypotheses (general and comparative intervention effectiveness) was conducted with a one-tailed α of .05. Specifically, designs with 2, 3, and 4 pairs of participants were examined based on 15 observations per participant. There were 5 potential start points for each pair, randomly selected from observations 6 through 10. For the general intervention effectiveness simulations, with single randomization each pair received the baseline phase (A) followed by the intervention (B) phase; in contrast, with dual randomization the pairs were randomly assigned to either an AB or BA order. For the comparative intervention effectiveness simulations, with single randomization the first pair member always received Intervention X and the second pair member Intervention Y; in contrast, with dual randomization, pair members were randomly assigned to the two intervention conditions.

The time-series data for each case were simulated as described in the previous investigations, with the standardized effect size for the pair member assigned to Intervention X set to d_1 and the standardized effect for the pair member assigned to Intervention Y set to d_2. For the general intervention effectiveness test, $d = (d_1 + d_2)/2$ was varied from 0 to 3 in increments of .5 by setting $d_1 = d_2 = d$. For the comparative intervention effectiveness test, $\mathrm{d} = d_2 - d_1$, d_1 was set to 0 and d_2 was varied from 0 to 3 in increments of .5. The latter effect size can be alternatively written as $d = [(\mu_{B2} - \mu_{A2}) - (\mu_{B1} - \mu_{A1})]/\sigma$, which is readily conceptualized and interpreted as a standardized ‘difference in differences’ (e.g., Marascuilo & Levin, 1970). The present measure differs from the standardized ‘half difference in differences’ effect-size estimator of $(d_2 - d_1)/2$ that is provided in Gafurov and Levin’s (2014) *ExPRT* program for the comparative intervention effectiveness hypothesis. The half difference-in-differences measure was incorporated into *ExPRT* because it represents a properly scaled interaction contrast when formulated for sample-size and power determination purposes from an analysis-of-variance perspective (Levin, 1997). It therefore should be kept in mind that a present power estimate associated with a difference-in-differences effect size of 2.00 corresponds to the power estimate associated with *ExPRT*’s half difference-in-differences effect size of 1.00.

Results

General intervention effectiveness hypothesis Dual- and single-randomization powers for Levin and Wampold's (1999) general intervention effectiveness hypothesis are presented in Figures 7 and 8 for autocorrelations of 0 and .3, respectively. The averaged pair power results presented in Figures 7 and 8 are easy to describe, especially when juxtaposed with Investigation 3a's individual results that were previously presented in Figures 3 and 4. Although the actual power values differ in the two investigations, the patterns involving single- and dual-randomization powers—namely, the magnitudes of the power advantage favoring the latter over the former—are remarkably similar. For example, when the total number of cases is held constant (e.g., 4 individuals in Investigation 3a, 2 pairs here; 6 individuals in Investigation 3a, 3 pairs here), with an autocorrelation of .3, mid-range effect-size values of $d = 1$ and 1.5, and two asymptotic power situations excluded, the six differences between the dual- and single-randomization powers all hover around .40. Specifically, from the graphs based on $N = 4$ individuals (Figure 4) and $N = 2$ pairs (Figure 8), it may be determined that the respective power differences are .43 and .37 for $d = 1$ and are .36 and .39 for $d = 1.5$; for $N = 6$ individuals and $N = 3$ pairs, the power differences are .42 and .40 for $d = 1$.

Comparative intervention effectiveness hypothesis Dual- and single-randomization powers associated with Levin and Wampold's (1999) comparative intervention effectiveness hypothesis are presented in Figures 9 and 10 for autocorrelations of 0 and .3, respectively. In each of those figures it may be seen that the dual-randomization procedure, which incorporates additional randomization-distribution outcomes as a result of randomly assigning pair members to the two interventions, X and Y, produces substantial power increases over Edgington's (1975) original single-intervention start-point procedure. For example, in Figure 10 based on an autocorrelation of .3, $N = 3$ pairs, and a difference-in-differences effect size of 2.0 (which corresponds to *ExPRT*'s half difference in differences of 1.0), power for the dual-randomization procedure is .87 as compared to only .46 for the single-randomization procedure.

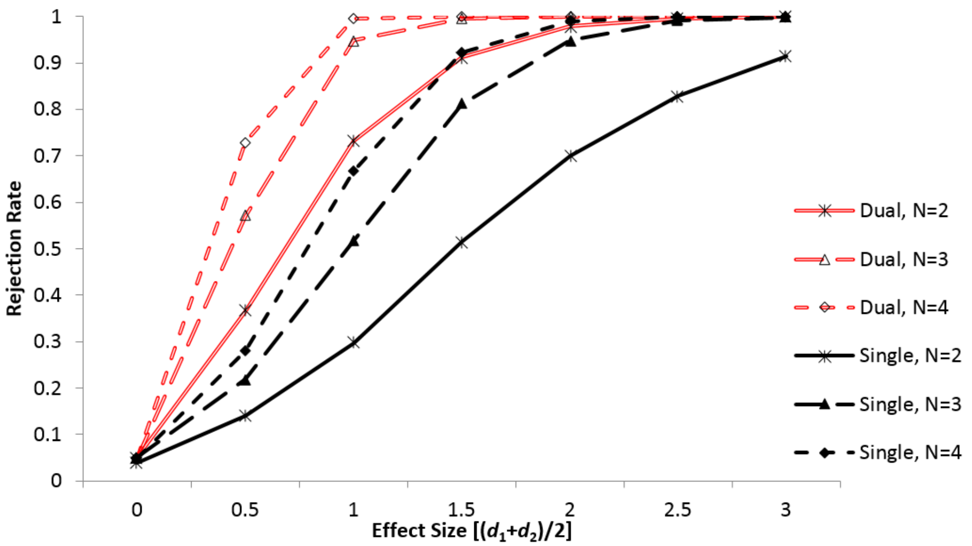

Figure 7

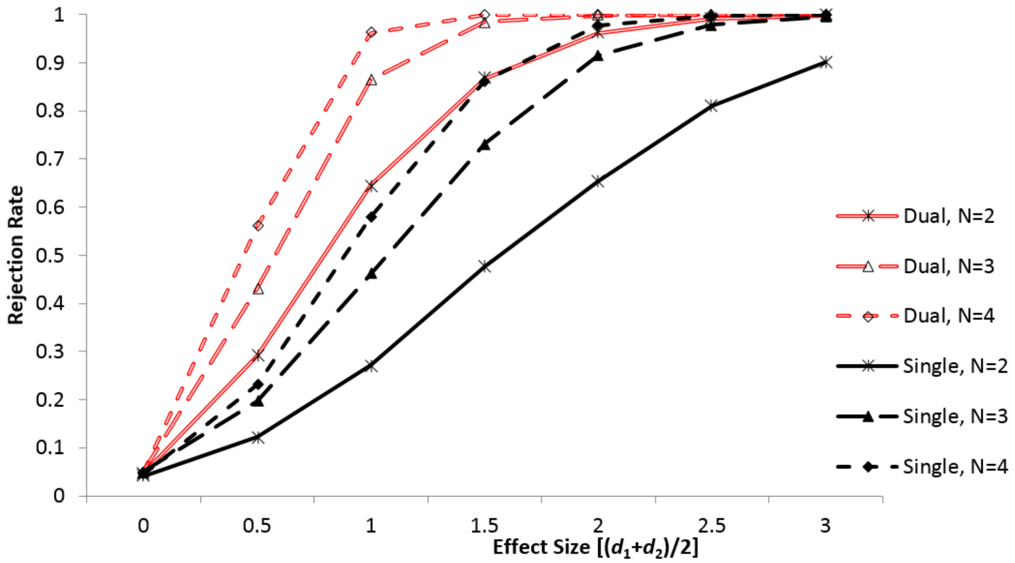

Figure 8

Figures 7 and 8. Investigation 4: Comparison (α = .05, one-tailed) of powers for the Single and Dual randomized general intervention effectiveness hypothesis replicated across *N* pairs. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of 0 (**Figure 7**) or .3 (**Figure 8**).

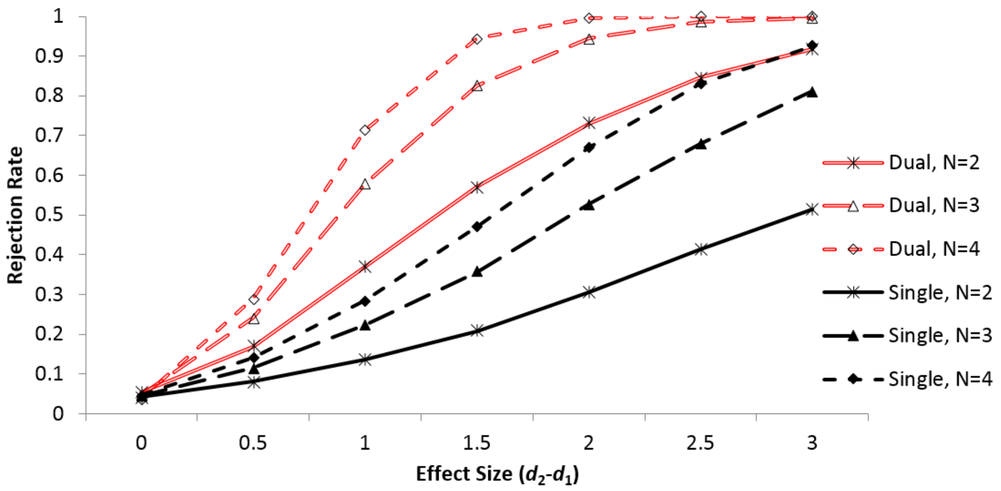

Figure 9

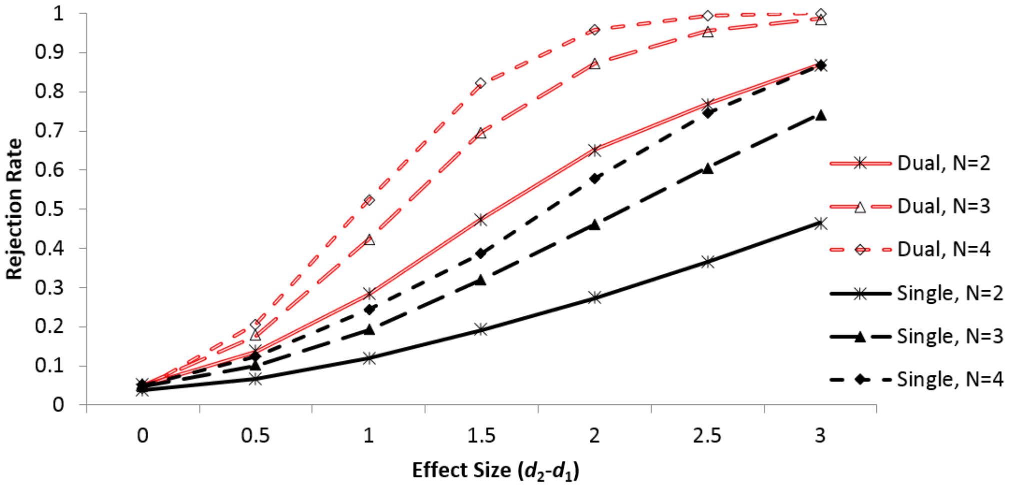

Figure 10

Figures 9 and 10. Investigation 4: Comparison (α = .05, one-tailed) of powers for the Single and Dual randomized Levin-Wampold comparative intervention effectiveness hypothesis replicated across *N* pairs. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of 0 (**Figure 9**) or .3 (**Figure 10**). Effect sizes are defined in a difference-in-differences metric, which correspond to half difference-in-differences effect sizes given by the present values divided by 2 (see text for further discussion).

Discussion

The present single-randomization powers associated with both the general and comparative intervention effectiveness hypotheses are quite consistent with those reported in an earlier simulation study by Lall and Levin (2004). However, the results of Investigation 4 make it clear that whenever either AB phase randomization is employed (general intervention effectiveness hypothesis, as is also manifested in Investigations 1-3) or the pair members are randomly assigned to the two intervention conditions, X and Y (comparative intervention effectiveness hypothesis), then the researcher can justifiably incorporate that randomization component into the randomization test. Doing so produces a large power boost relative to Levin and Wampold's (1999) original randomization tests that incorporate only intervention start-point randomization. The impressive dual-randomization power increases for the comparative intervention hypothesis are particularly noteworthy and heretofore undocumented. Although Levin and Wampold recognized the methodological (internal validity) necessity of randomly assigning the XY pair members to intervention conditions when testing that hypothesis, their single-randomization test procedure does not capitalize on the statistical power benefits that result from random assignment.

At the same time, and as was suggested by Levin and Wampold (1999, p. 78), now suppose that instead of X and Y representing two alternative interventions to which pair members are randomly assigned (as was examined here), they represent some non-randomly assigned participant-differentiating (or *status*) variable of interest (e.g., gender, age, ability, amount of prior experience), where one pair member (X) represents one level of the status variable (e.g., male, older, higher, more prior experience) and the other pair member (Y) represents a different level (female, younger, lower, less prior experience). In that nonrandom-assignment situation, the additional 2^N X vs. Y randomization outcomes of the modified Levin-Wampold formula (provided earlier in this section) cannot be incorporated into the randomization distribution, in which case the statistical test would revert to the original procedure developed by Levin and Wampold. It should be noted, however, that: (1) the inclusion of the status variable (e.g., gender, age, ability, amount of prior experience) still permits the investigation of a possible intervention-by-status interaction (e.g., the intervention is relatively more effective for individuals with less prior experience than for individuals with more prior experience) with the comparative intervention effectiveness test; and (2) if AB phase randomization is included in a nonrandomized status-variable study, then the 2^N factor associated with phase randomization in the modified Levin-Wampold general intervention effectiveness formula (provided earlier in this section and the primary focus of the present order-randomization study) would reappear.

Let us additionally consider a participant-pairs situation in which both the XY and the AB factors have randomized components. For example, X and Y could represent two randomly assigned instructional interventions: experimental vs. control (as in Investigation 4, and the primary factor of interest); and A and B could represent two types of practice: teacher- vs. self-directed (the secondary factor of interest), the order of which is randomly assigned to each pair. In that situation, the currently investigated two-factor randomization design (intervention start points and phase orders) could be expanded to encompass a third randomized factor (intervention start points, instructional intervention, and practice-type phase order). Yet, it is important to note that: (1) incorporating *either* AB or XY randomization into the Levin-Wampold (1999) simultaneous pairs design will enhance the design's internal validity and produce a statistical power increase to detect general (AB) or comparative (XY) intervention effectiveness, relative to the power of the original procedure; and (2) although incorporating *both* AB and XY randomization components into the design (as in the present three randomized factor design example) provides a double internal-validity enhancement, the resulting power is exactly the same as that associated with incorporating only one of these additional randomization components (i.e., either AB or XY).

Investigation 5: Randomized Intervention Order for the Single-Case Crossover AB Design

The crossover design is a standard investigative strategy in conventional-group educational intervention research (see, for example, Jones & Hall, 1982; and Levin et al., 1990, Exp. 1). With a crossover design it is possible to compare two intervention conditions (or an intervention and a nonintervention control condition) in two independent groups that also receive both intervention conditions in counterbalanced orders. Although various single-case designs (e.g., the alternating treatment design) allow for each case to receive two or more interventions, the within-case structuring and/or rapid alternation of treatments does not provide an adequate parallel to capture the essence of the crossover design. With a little tweaking, however, the present order-randomization approach can be adapted to capture that essence.

With A and B representing two different interventions, the present order-randomization modification of Marascuilo and Busk's (1988) model has all the apparent trappings of a crossover design. However, adding a straightforward order-randomization component to that model may not adequately fit a single-case researcher’s crossover-design bill. Specifically, randomizing the intervention order independently for all participants (or other units) in the Marascuilo-Busk model does not guarantee that an equal number of participants will receive the two orders, AB and BA—something that is desirable, if not essential, for producing a study that is completely counterbalanced with respect to the order of intervention administration. In fact, in the extreme, a simple randomization scheme could actually result in all participants receiving the same order of intervention administration. In a single-case intervention study with a small number of cases, that situation is not as unlikely as it may initially appear. For example, with $N = 2$ cases it will happen half the time; with $N = 3$ it will happen 25% of the time; and with $N = 4$, it has a 12½% chance of occurring. It should also be recognized that it is not possible to have *complete* (i.e., perfect) order counterbalancing with an odd number of participants.

Consequently, a potentially useful alternative is a crossover design that is completely counterbalanced with respect to the order in which the two different interventions are administered. Implementing such a procedure perfectly controls for potential contaminating effects associated with the two different intervention orders (AB and BA) and therefore eliminates order effects as an internal validity concern. This can be accomplished with a *restricted randomization* scheme, the Type I error and power characteristics of which are explored next in the context of Investigation 5.

Method

In this investigation we examined the effect on Type I error and power characteristics of restricting the dual-randomization scheme to ensure a balance between cases assigned to crossover design orders AB and BA. Specifically, a restricted dual-randomization crossover design (henceforth referred to as *restricted*) with 15 observations and $k = 5$ potential start points for each case randomly selected from observations 6 through 10 was examined for conditions with 2, 3, 4, 5, and 6 cases. For conditions with an even number of participants the number assigned to AB was restricted to equal the number assigned to BA, resulting in a augmented multiplier factor of $\binom{N}{x} = N!/x!(N-x)!$ to the k^N potential intervention start-point randomization outcomes (or $\prod_{i=1}^{N} k_i$ when the number of potential intervention start points differs across cases), where N is the total number of cases and x is the number of cases that are to be randomly assigned to each of the two administration orders. For an odd number of participants the number assigned to AB was restricted to equal the number assigned to BA, plus or minus 1. In the latter (odd number) case, because of the dual-randomization process of: (1) randomly determining which order, AB or BA, was to be associated with the larger number; and (2) randomly assigning the two orders to participants, this resulted in an augmentation factor of $2\binom{N}{x} = 2\left[N!/x!(N-x)!\right]$ (see Levin et al., 2014, p. 192). Effect sizes were varied from 0 to 3 in increments of .5, again the autocorrelation was set to 0 or .3, and one-tailed α = .05 tests were conducted.

Results

Results from the conditions where the autocorrelation is 0 are shown in Figure 11, whereas those for an autocorrelation of .3 are shown in Figure 12. For comparative purposes, results from the unrestricted-dual randomization designs (henceforth referred to as *unrestricted*) of Investigation 3a are also included in those two figures. In Figures 11 and 12 it is clear that for all sample sizes the restricted-randomization tests yielded empirical Type I errors (i.e., when the effect size was 0) that corresponded with their nominal .05 values. Although it is evident from Figures 11 and 12 that the restricted-randomization crossover-design powers are uniformly lower than the corresponding unrestricted-randomization powers, the difference between the two becomes less and less noticeable with increases in sample size. With *N*s of 5 and 6, for example, the power differences are negligible for all practical purposes. At the same time, it should be pointed out that even at the smaller sample sizes the restricted-randomization crossover-design powers are respectable. To wit, in Investigation 3a it was indicated that with an autocorrelation of .3 and $N = 3$ participants, the unrestricted-randomization test's power for detecting an effect size of $d = 1.5$ was equal to .90 (reproduced in Figure 12); and as may also be seen in Figure 12, for the same set of parameters the restricted-randomization crossover-design test's power is .865.

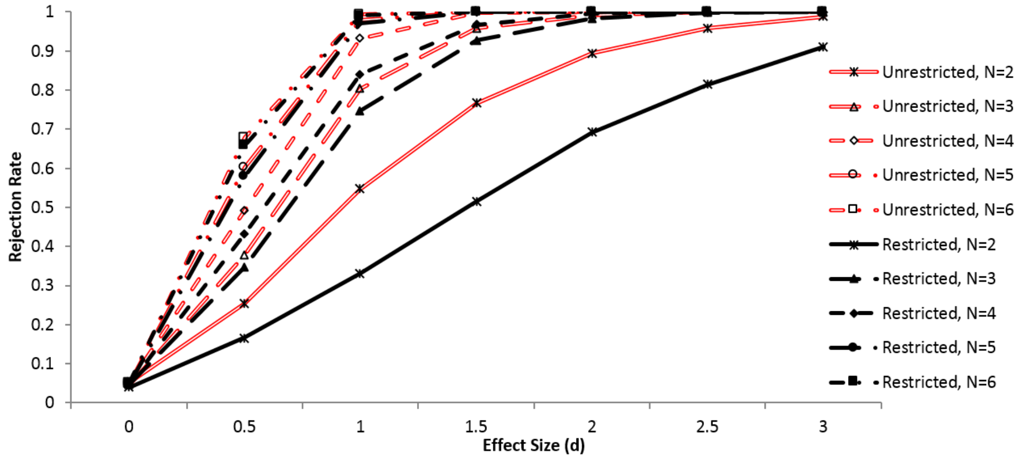

Figure 11. Investigation 5: Comparison (α = .05, one-tailed) of randomization tests for the Restricted Dual and Unrestricted Dual AB randomized crossover designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of 0.

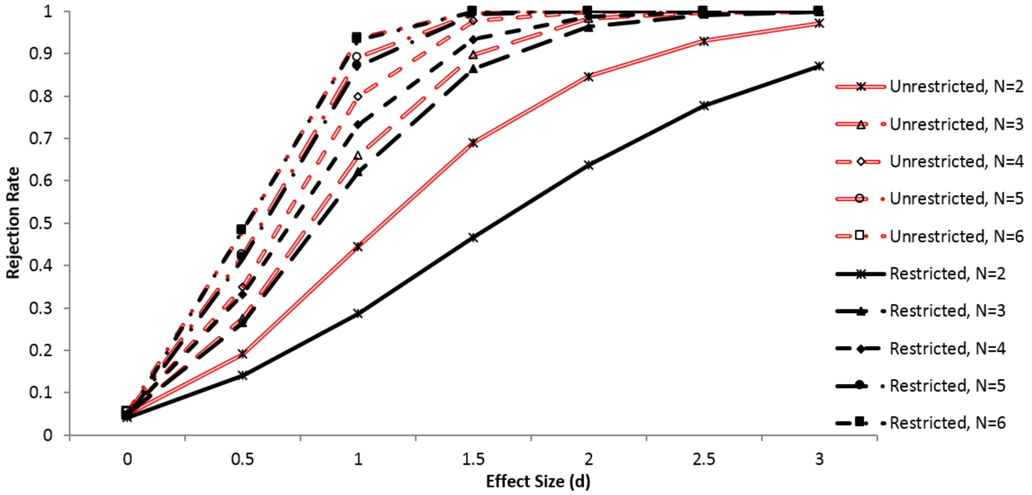

Figure 12. Investigation 5: Comparison (α = .05, one-tailed) of randomization tests for the Restricted Dual and Unrestricted Dual AB randomized crossover designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for a 15 observations design with 5 potential intervention start points designated from between the 6th and 10th observations inclusive and an autocorrelation of .3.

Investigation 6: Randomized Intervention Order for the Single-Case ABAB Design

In Investigation 5 the simulations were extended to four-phase ABAB designs (also referred to as *reversal* or *operant* designs—see, for example, Kratochwill & Levin, 2010). More specifically, Type I error and power were examined for Onghena's (1992) randomized intervention start-point ABAB design (Single) and a combined randomized intervention start-point plus random-order (ABAB versus BABA) design (Dual), with the dual approach enhancing the ABAB design's internal validity by virtue of its controlling for potentially confounding order effects.

Method

The effect of case replications (more participants) on power was examined for a design with 23 observations and a minimum of 5 observations in each of the four phases, which implies that the number of possible permutations for one case is 20 for the single-randomized design (for computational details, see Onghena, 1992) and 40 for the dual-randomized design. The simulations included 1, 2, 3, or 4 participants, effect sizes that varied from 0 to 3 in increments of .5, and an autocorrelation of 0 or .3. Sample sizes greater than 4 were not investigated because ABAB designs provide more intervention-effect information per case than AB designs and thus they tend to be replicated across fewer participants. Thus, the value in extending the study to larger numbers of participants was judged not to warrant the increased computational time that would have been required. All tests (based on the average of the two B-phase observations minus the average of the two A-phase observations) were conducted with a one-tailed Type I error probability of .05. In that regard, it should be mentioned that the present simulations are based on the weighted (by the number of outcome observations, O) A- and B-phase means [i.e., $(\mathrm{O}_{\mathrm{A1}}M_{\mathrm{A1}} + \mathrm{O}_{\mathrm{A2}}M_{\mathrm{A2}})/(\mathrm{O}_{\mathrm{A1}} + \mathrm{O}_{\mathrm{A2}})$ and $(\mathrm{O}_{\mathrm{B1}}M_{\mathrm{B1}} + \mathrm{O}_{\mathrm{B2}}M_{\mathrm{B2}})/(\mathrm{O}_{\mathrm{B1}} + \mathrm{O}_{\mathrm{B2}})$] whereas Gafurov and Levin's (2014) *ExPRT* program calculations are based on the unweighted means [$(M_{\mathrm{A1}} + M_{\mathrm{A2}})/2$ and $(M_{\mathrm{B1}} + M_{\mathrm{B2}})/2$]. Power differences attributable to the two weighting schemes *per se* should be minimal for the set of parameters that were specified for the present simulations, however.

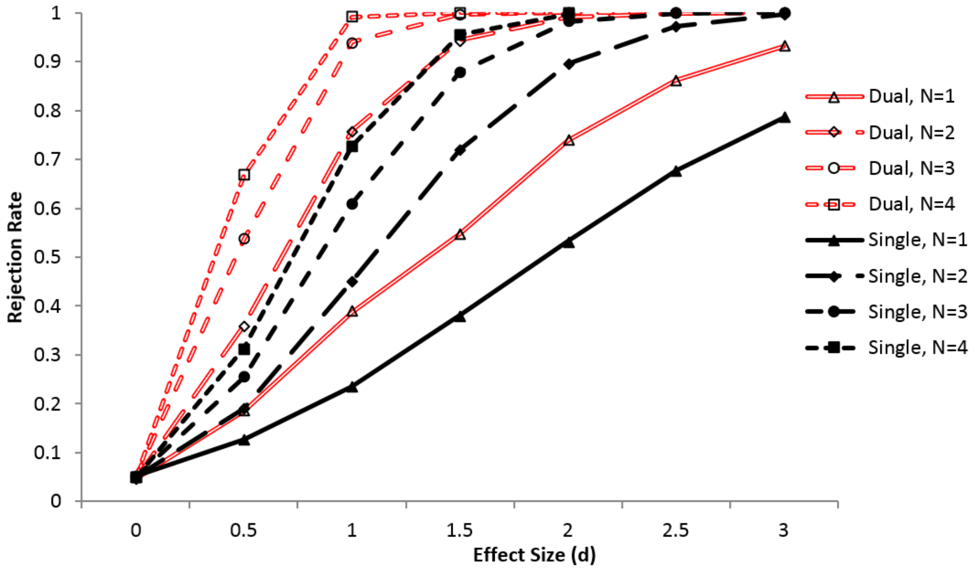

Figure 13

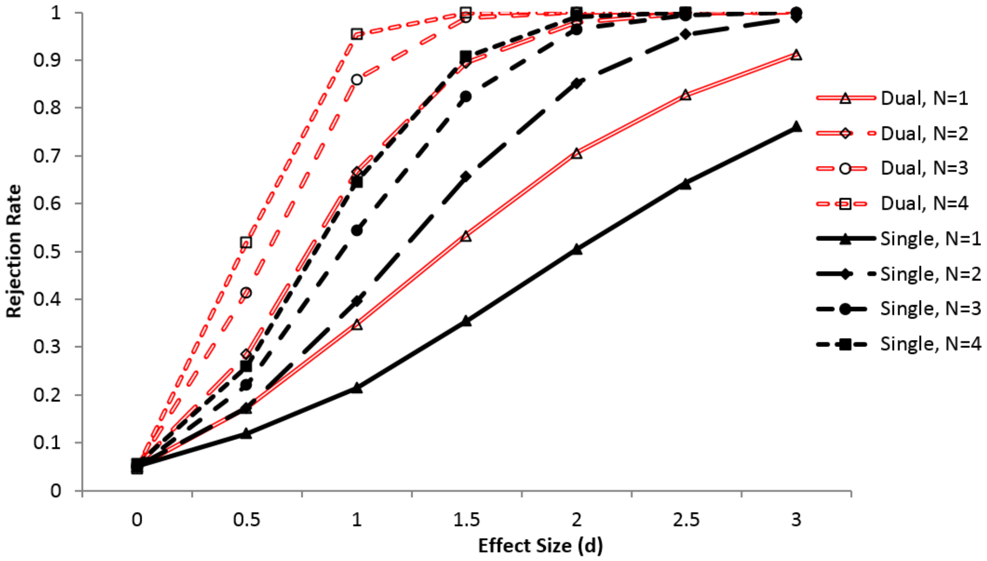

Figure 14

Figures 13 and 14. Investigation 6: Comparison (α = .05, one-tailed) of randomization tests for the Single and Dual randomized ABAB designs replicated across *N* cases. The rejection rate of the null hypothesis is shown as a function of effect size and *N*, for an autocorrelation of 0 (**Figure 13**) or .3 (**Figure 14**), and a 23 observations design with a minimum of 5 observations in each of the four phases. The resulting number of possible randomizations is 20 for the Single randomization scheme and 40 for the Dual randomization scheme.

Results

Results from the conditions where autocorrelation was 0 are shown in Figure 13, while those for an autocorrelation of .3 are shown in Figure 14. As was true for the AB designs, once again the present dual-randomization scheme greatly overpowers the single-randomization scheme. For example, with an autocorrelation of .30 an effect size given by d = 1.5, and an N = 2 design, single-randomization ABAB power is equal to .66 whereas dual-randomization ABAB power is .895—a nontrivial power difference of almost .24. For the single-randomization scheme to achieve comparable power (.91) to that of the dual-randomization scheme (.895) would require twice as many participants, namely N = 4.

Investigations 7 and 8: The Single-Case AB Design Revisited

What follow are two additional AB design investigations, both of which follow directly from colleagues' concerns about data characteristics of the simulations reported thus far. One such concern focuses on the series lengths associated with all of the simulations conducted so far and the other focuses on the distributional characteristics of the outcome measure that comprises all of those simulations. These two concerns are addressed in Investigations 7 and 8, respectively.

Investigation 7

In a recent survey of single-case intervention research reported in 21 journals and based on 809 cases during the year 2008, Shadish and Sullivan (2011) reported that the modal and median series length per case consisted of 20 total observations. The positively skewed distribution had a mean of 27.0 and range of 2 to 160. Approximating from Shadish and Sullivan's frequency histogram (Figure 2), one can estimate that 23% of the cases had series lengths in the 20-29 range, with 16% in the 30-39 range, 6% in the 40-49 range, and 5% that were 50 or more. Moreover, it is not difficult to locate single-case intervention studies in recent years that included 50 or more outcome observations per case—see, for example, Lucynski, Hanley, & Rodriguez (2014), with 6 children and approximately 50 observations per child; Pellecchia et al. (2011), with 8 children and 60 or more observations per child; Hanley, Jin, Vanselow, & Hanratty (2014), with 3 children and approximately 70 observations per child; and Donaldson, Trahan, & Kahng (2014), with 1 adult exhibiting dementia and approximately 130 observations.

In the present Investigation 1, the simulation consisted of 30 outcome observations; in Investigation 2, the range spanned from 20 to 150; in Investigations 3 and 4 there were 15 outcome observations; in Investigation 5 there were 15 and 30; and in Investigation 6 there were 23. Therefore, the series lengths for the present simulations do not seem too far out of line with those of single-case intervention studies that are being reported in the literature, where at least half of them include at least 20 observations (Shadish & Sullivan, 2011).

Why, in the first place, was a series as long as 30 decided upon for our Investigations 1 and 5? The answer is simple with respect to the primary focus of the study. Specifically, at least 21 observations (i.e., 20 potential intervention points with at least one baseline observation and one intervention observation) are required to compare Edgington's (1975) single randomization-test procedure and the present dual modification based on a one-tailed α of .05. We settled on 30 total observations to provide at least 5 baseline observations and 5 intervention observations, thereby obtaining some degree of stability in those two series.

That said, in Investigation 7 we examined whether the already reported power difference favoring the dual- over the single-randomization approach would generalize to shorter—in fact, very short—series ($N < 10$), as was analogously examined by Levin et al. (2011) in their short series Investigation 2's AB design.

Method Here, the simulation parameters and procedures of Investigation 3 were again selected and applied to three short-series conditions. Power for each of these conditions was assessed for the single- and dual-randomization test procedures (α = .05, one-tailed) for both series based on an autocorrelation of 0 and those based on an autocorrelation of .30.

In one condition two cases were included, with 9 outcome observations per case. The first two observations were always in the first phase, the last two observations were always in the last phase, and the intervention start point was randomly chosen from among the middle five observations in the series. In a second condition three cases were included, with 7 outcome observations per case. The first two observations were always in the first phase, the last two observations were always in the last phase, and the intervention start point was randomly chosen from among the middle three observations in the series. The third condition consisted of five cases, with 8 outcome observations per case. The first three observations were always in the first phase, the last three observations were always in the last phase, and the intervention start point was randomly chosen from among the middle two observations in the series.

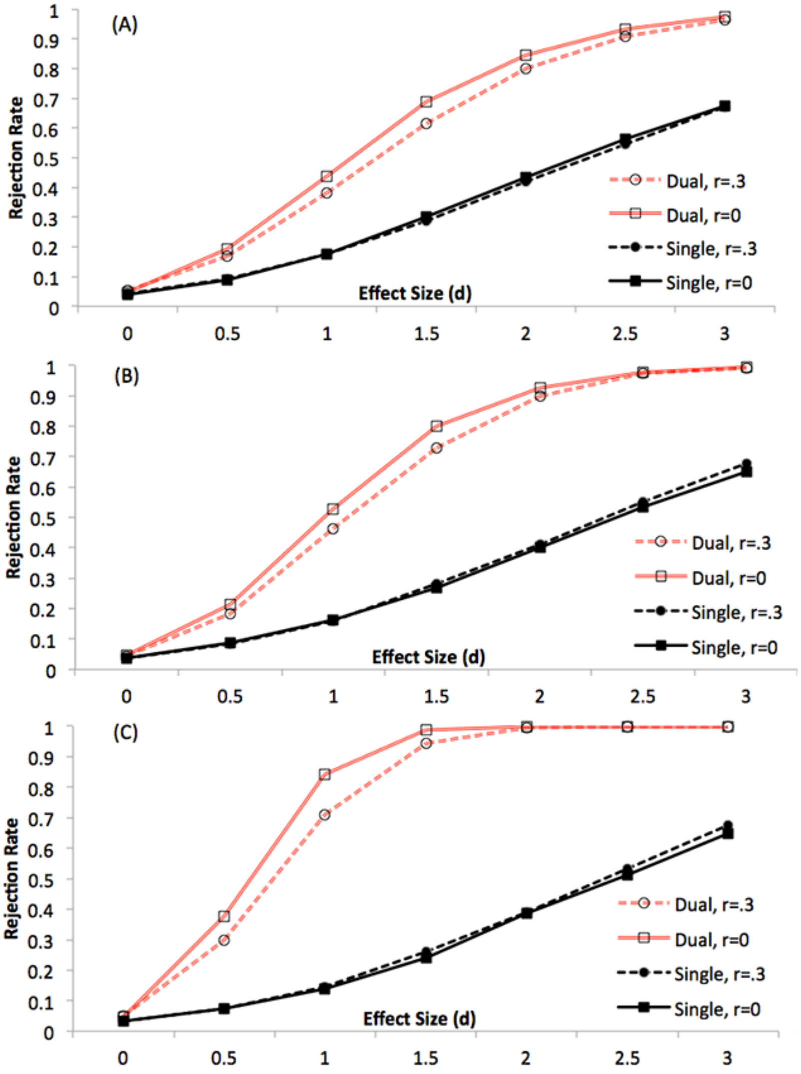

Figure 15. Investigation 7: Comparison (α = .05, one-tailed) of randomization tests for the basic AB randomized intervention start-point design (Single) and the randomized intervention start-point plus randomized intervention-order design (Dual). The rejection rate of the null hypothesis is shown as a function of the effect size and level of autocorrelation for: (A) a two-participant design with nine observations each where the start point is randomly assigned to one of the middle five observations, (B) a three-participant design with seven observations each where the intervention start point is randomly assigned to one of the middle three observations, and (C) a five-participant design with eight observations each where the start point is randomly assigned to one of the middle two observations.

Results The results are summarized in the three panels of Figure 15, where it may clearly be seen that, as in Investigation 3a, with the Type I error well controlled, in all three conditions the dual- randomization test's powers by far surpass those of the single-randomization test. A direct comparison of selected dual-over-single power advantages in the long-series Investigation 3a (Figure 4) and the present short-series investigations (Figure 15) is summarized in Table 3, where it should be noted that the advantages in the short-series investigations are comparable to (or larger than) those of the long-series investigations. On that basis, it can be concluded that the appeal of the dual-randomization approach is not restricted to long-series intervention studies. The approach applies equally well, if not better, to intervention studies consisting of a total of 7, 8 or 9 outcome observations.

Table 3. Selected single- versus dual-randomization power comparisons of the present longer (Investigation 3a, Figure 4) and shorter (Investigation 7, Figure 15) series simulations (SL = Series Length, PISP = Number of Potential Intervention Start Points)

N	***d***	***r***	**Size (SL/PISP)**	**Single**	**Dual**	**Difference**
2	2	0.3	Longer (15/5)	0.44	0.85	0.41
			Shorter (9/5)	0.42	0.8	0.38
3	1.5	0.3	Longer (15/5)	0.49	0.9	0.41
			Shorter (7/3)	0.28	0.73	0.45
5	1	0.3	Longer (15/5)	0.45	0.89	0.44
			Shorter (8/2)	0.15	0.71	0.56

As may also be seen in Figure 15, in contrast to the long-series results presented in Figures 3 and 4, throughout the present study, and in previous investigations, the powers associated with the single-randomization approach do not decrease as the autocorrelation increases from 0 to .30. In fact, a slight power increase may be observed for the larger effect sizes in Panels B and C. This same positive relationship between autocorrelation and power for the single-randomization approach was also discovered and noted by Levin et al. (2011) in their short-series Investigation 2. Those authors offered a speculative interpretation of that finding, but a experimental examination of that interpretation remains to be conducted.

Investigation 8

In all of the present simulations, the data were generated assuming that the outcome measure was continuous and normally distributed, whereas in many single-case intervention studies the outcome measures consist of discrete counts or rates. Therefore, to assess whether the power differences favoring the dual-over-single randomization approach would be observed even in an extremely non-normal distribution situation, Investigation 1 was replicated with the only change being that the outcome measure was simulated to be a binary variable as opposed to a continuous one.

Method More specifically, the same algorithms were used to generate the data, but the resulting values were dichotomized such that all values over 1 were recoded as 1 and all values under 1 were recoded as 0. Thus, for conditions without autocorrelation, the baseline observations had a probability of .34 of being a 1 (and .66 of being a 0), whereas the probability of obtaining a 1 in the intervention phase depended on *d* (e.g., when *d* equaled 0, 1, 2, 3, 4, and 5, the probabilities of obtaining a 1 were .34, .50, .84, .98, .999, .99997).

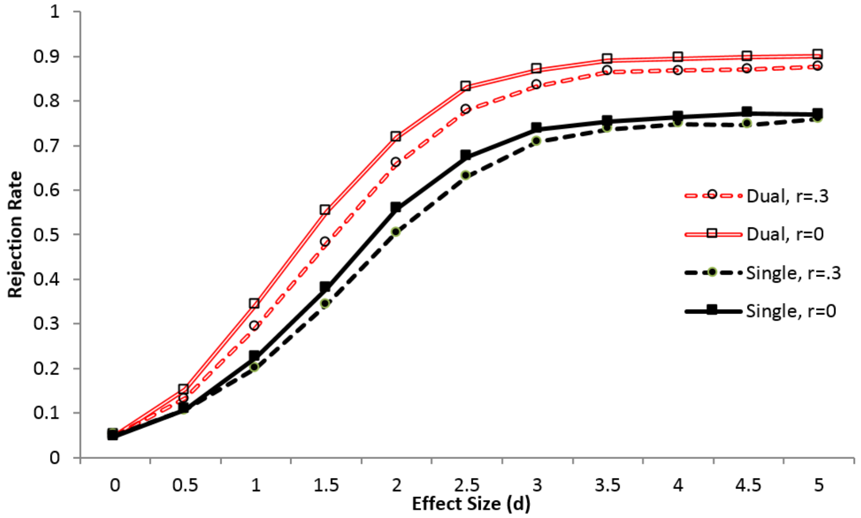

Figure 16. Investigation 8: Comparison (α = .05, one-tailed) of randomization tests for the basic AB randomized intervention start-point design (Single) and the randomized intervention start-point plus randomized intervention-order design (Dual), where the outcome is binary and the intervention start point is randomly selected between the 6th through the 25th observations inclusive in a 30-observations study. The rejection rate of the null hypothesis is shown as a function of the effect size and level of autocorrelation.

Results The results of this simulation may be seen in Figure 16. Similar to when the continuous outcome was examined (Investigation 1) the dual-randomization approach consistently leads to greater power than the single-randomization approach, but as would be anticipated, the dichotomization of the outcome lessens the power for each. Also of note, the power estimates reach a ceiling below 1.0, which can be explained by the baseline observations being set so there was a .34 probability of observing the desired behavior. If the baseline probability had been set lower, say to .01, the difference in probabilities between phases could be larger, leading to higher observed maximum powers.

General Discussion

In the eight Monte Carlo investigations reported here, we discovered that in situations where researchers are able to randomize the order in which the phases of single-case AB and ABAB designs (or the interventions themselves in paired-cases designs) can be administered by, for example, simple coin flips, it is clearly advantageous to do so. Order randomization represents a valuable addition to Edgington's (1975) and Onghena's (1992) randomized start-point models, in that it: (1) enhances those designs' internal validity (a methodological improvement); and (2) effectively controls the associated randomization test's Type I error probability, while affording increases in the test's power (a statistical improvement). In many of the instances examined, these power increases were dramatic with respect to a single-case researcher's economic savings. For instance, in Investigation 2's $N = 1$ simulations we found that an AB design with the present dual-randomization scheme could require less than half as many outcome observations as Edgington's original single-randomization scheme. Specifically, as may be seen in Figure 2, for $\alpha = .05$ (one-tailed), an effect size of 2.0, and a series autocorrelation of .3, the dual-randomization approach based on 30 outcomes yields power of .67. In contrast, to achieve similar power with the single-randomization approach requires between 80 and 90 outcome observations. In alternative economic terms, in Investigations 3 and 5 we found that in $N > 1$ investigations, about twice as many participants are required for the single-randomization approach to achieve power equivalent to that of the dual-randomization approach (see Figures 3-8). Similar dual-over-single randomization power advantages were achieved in the Investigations 4 and 6 randomized pair-members AB design and four-phase ABAB design, respectively. Importantly to single-case researchers from both practical and versatility perspectives, such power advantages were also observed in: (a) short-series designs consisting of as few as seven observations (Investigation 7); and (b) single-case intervention contexts associated with binary, rather than normally distributed, outcome measures (Investigation 8).

Additional Considerations for the Single-Case Crossover Design

Restricted or unrestricted randomization: Which is better? To guarantee order balance (and, therefore, greater internal validity) in single-case AB crossover designs, a restricted dual-randomization scheme must be employed, rather than an unrestricted one. Although the restricted-randomization approach results in powers that are uniformly lower than those associated with an unrestricted-randomization approach, as sample sizes increase beyond $N = 2$ or 3 cases the respective powers of the two designs are quite comparable. So, whenever a researcher is considering the tradeoff between a guaranteed crossover-design balance of intervention administration order (thereby controlling perfectly for order effects), on the one hand, and some degree of increased statistical power, on the other, then: (1) if the former is considered to be relatively more important, the researcher should select the restricted-randomization procedures of Investigation 5; and (2) if the latter wins out as being relatively more important, the researcher should choose the unrestricted-randomization procedures of Investigation 3, especially when the sample size is relatively small (i.e., $N < 3$ or 4 cases).

Controlling for potential confounding factors In actual intervention research studies based on within-subjects designs, in general, and single-case AB crossover designs, in particular (as represented by current Investigation 5), more potentially confounding variables than simple order effects must be taken into account and controlled. That is, between-phase outcome changes may also be the result of other extraneous factors, including: external effects, such as those attributable to history; effects associated with the experimenter or instructor; and effects associated with the participant, such as novelty, Hawthorne, and "John Henry" effects (see, for example, Shadish, Cook, & Campbell, 2002). Such confounding variables can severely compromise an intervention study's internal validity—namely, that the manipulated intervention *per se* was responsible for between-phase outcome changes—as well as its construct validity. In research now in progress, we are comparing the effects of extraneous factors on internal and statistical-conclusion validity in the present unrestricted and restricted crossover designs.

A random-assignment caveat A few words of operational caution connected to the restricted design crossover design of Investigation 5 should be offered to interventionists who elect to implement that design in their research. Specifically, some researchers are likely to make a critical random-assignment mistake when it comes to implementing the randomization process correctly. With an even number of cases, there should be no problem, in that the researcher would randomly select half of the cases to receive an AB order of intervention administration, with the remaining half receiving the BA order. With an odd number of cases, however, the researcher needs to consider possible assignments where either the AB order or the BA order receives the larger number of cases. To do so, the researcher could go through a two-step randomization process, as follows. In Step 1, the researcher would randomly determine whether the larger number of cases is to receive the AB order or the BA order (e.g., 4 cases if $N = 7$). Then in Step 2, the researcher would proceed as in the previous "even N" situation, namely randomly selecting the N_1 cases that will be receiving the AB order, with the remaining N_2 cases receiving the BA order. Without the researcher conducting the restricted-randomization procedure in this two-step fashion (or through an analogous completely random-assignment process), subjectivity would enter into the researcher's decision about which order (AB or BA) receives the one more (or one fewer) case, resulting in the randomization distribution and its associated statistical test being invalid.

Levin and Wampold's (1999) Simultaneous Pairs Intervention Start-Point Model Revisited

In the present Investigation 4, we examined Levin and Wampold's (1999) simultaneous pairs, comparative intervention effectiveness hypothesis, with a randomized XY intervention variable included in the randomization-test analysis. In that situation, we found the statistical power of the procedure to be greatly enhanced relative to that of the original Levin-Wampold procedure, for which the randomized intervention factor is not taken into account. We now consider a variation and an extension in conjunction with the present modified procedure.

For the variation, suppose that the A and B phases represent two competing interventions and, as in Investigations 1-3, it is possible to randomize the order in which the two phases are administered (A followed by B or B followed by A). Within each participant pair, it is randomly determined which pair member is assigned the AB administration order and which the BA order (say, X = AB and Y = BA). The data are collected and, as in Investigation 4, the comparative intervention effectiveness hypothesis is tested (with the inclusion of the 2^N multiplier associated with the randomized XY factor) on the difference in differences, $(X_{A1} – Y_{B1}) – (X_{B2} – Y_{A2}) = (X_{A1} + Y_{A2}) – (X_{B2} + Y_{B1})$. Note that in this context the interaction actually represents a main effect comparison of Intervention A vs. Intervention B, just as it does in a conventional crossover design. Accordingly, this paired-cases design then becomes conceptually equivalent to the just discussed restricted-order crossover design of Investigation 5, but because of the pairs structure here, for which it is guaranteed that: (1) there will be equal numbers of participants receiving each intervention order; and (2) within each pair, the crossover will occur at exactly the same point in time.

For the extension of the modified Levin-Wampold (1999) simultaneous pairs comparative intervention effectiveness test, now suppose that two equivalently scaled (or commensurable) outcome measures, M_1 and M_2, are constructed to be differentially sensitive to an intervention; or alternatively, that M_1 is expected to be more responsive to Intervention X than to Intervention Y and M_2 is expected to be more responsive to Intervention Y than to Intervention X—as with Levin's (1989) experimental illustrations of Campbell and Fiske's (1959) discriminant validity and Morris, Bransford, and Franks' (1977) transfer-appropriate processing. The modified dual-randomization procedure to test Levin and Wampold's comparative intervention effectiveness hypothesis can be readily extended to accommodate thedifferential outcome-measure effects addition. Specifically, with X and Y representing randomly assigned interventions within each pair, A and B representing baseline and intervention phases (as in Investigation 4), and M_1 and M_2 representing commensurable measures or tests, the data to be analyzed are simply the intervention-by-phase difference-in-differences effect associated with M_1 minus the same effect associated with M_2, and which amounts to the three-way interaction of intervention by phase by outcome measure. This translates into an assessment of whether whatever differential change from Phase A to Phase B that is produced by the two interventions is the same on the two outcome measures. As with the Investigation 4 test of the two-way intervention-by-phase interaction (i.e., the comparative intervention effectiveness test), the statistical power to test this extended difference would also benefit from the 2N multiplier resulting from within-pair randomization of the intervention factor.

Extensions to Other Single-Case Intervention Designs and Situations

Other single-case designs Research by the present authors is currently in progress to extend the present randomized-order design-and-analysis procedure (combined with randomized intervention start points) to single-case intervention designs other than the AB-type and ABAB designs that were investigated here. Our initial efforts have been targeted at alternating treatment designs (Levin et al., 2012) and multiple-baseline designs. In the former, independently randomizing the alternating A and B intervention phases both within and across participants has been recommended as an internal-validity enhancer (e.g., Kratochwill & Levin, 2010) and incorporating both randomized intervention start points and randomization statistical tests into the process is relatively straightforward. In the latter, although multiple-baseline designs typically include a set of staggered baseline (A) and intervention (B) phases across participants, the present randomized-order approach could be adopted for situations in which, as was discussed here, an initial mandatory A' series of baseline (warm-up or adaptation) observations is included. The approach might also be possible in situations where A represents a standard or basic instructional/behavioral practice and B represents a competing alternative practice.

Other outcome measures As well as testing for between-phase mean (level) changes, the present randomized-order procedure is similarly applicable to testing for changes in slope (trend) and variance (variability). All such tests are available in Gafurov and Levin's (2014) Excel$^{©}$-based randomization-test software, which is freely accessible from the Google Drive *ExPRT* (Version 1.2) website, https://code.google.com/p/exprt/. At the same time, simulation research now in progress (Levin et al., 2014) is assessing the Type I error probabilities and statistical powers of the present combined randomized intervention start-point and randomized-order approaches relative to Koehler and Levin's (1998) randomized intervention start-point approach alone, with respect to tests of slope and variance, in various single-case intervention designs.

Other intervention effect types It is important to note that in the present eight-investigation set of Monte Carlo simulations, all intervention effects were modeled to represent immediate abrupt changes in the participant's mean level: that is, a constant increase in the participant's series of observations that is coincident with the initial potential intervention point specified by the researcher—or, in the case of the four-phase ABAB design, coincident with the initial potential phase-change (transition) point that was specified for each of the three phase changes. In some of our research in progress we are modeling other types of intervention effects as well, such as immediate gradual effects, delayed abrupt effects, and delayed gradual effects (see, for example, Lall & Levin, 2004). In each of these ongoing simulation studies our goal is to determine whether the present randomized-order approach and associated randomization test afford power benefits that are as impressive in other single-case design contexts (and for other outcome measures) as were discovered in the present AB and ABAB design tests of between-phase changes in level.

Final Comments

Although randomization schemes of the type advocated here may be opposed by single-case intervention researchers who have been steeped in the response-guided tradition (see, for example, Ferron & Levin, 2014), we hope that such schemes will be received more positively by at least some traditional single-case interventionists. In fact, for years many alternating-treatment design users have been diligent in assigning interventions to phases or sessions using a block-randomization process (Kratochwill & Levin, 2010; for a research example, see Holden, Bearison, Rode, Kapiloff, Rosenberg, & Rosenzweig, 2002). As a cause for further optimism, an increasing number of single-case investigations that have incorporated various forms of randomization design and analysis are appearing in both student dissertations and the published literature (e.g., Ainsworth, 2014; Bardon, Dona, & Symons, 2008; Bice-Urbach, 2015; Bonnet, 2012; Lojkovic, 2014; Regan, Mastropieri, & Scruggs, 2005).

Acknowledgments

We are grateful to Anya Evmenova for providing us with useful research references. Correspondence concerning this article should be addressed to Joel R. Levin at *jrlevin@u.arizona.edu*.

References

Ainsworth, M. K. (2014). *Effectiveness of the ALL Curriculum to teach basic literacy skills to groups of students with severe disabilities and complex communication needs*. Unpublished doctoral dissertation, George Mason University, Fairfax, VA.

Auerbach, C., & Zeitlin, W. (2014). *SSD for R: An R package for analyzing single-subject data*. New York: Oxford University Press.

Bardon, L. A., Dona, D. P., & Symons, F. J. (2008). Extending classwide social skills interventions to at-risk minority students: A preliminary application of randomization tests combined with single-subject methodology. *Behavioral Disorders*, *33*, 141-152.

Bice-Urbach, B. (2015). *Teleconsultation: The use of technology to improve evidence-based practices in rural communities.* Unpublished doctoral dissertation, University of Wisconsin-Madison, Madison, WI.

Billette, V., Guay, S., & Marchand, A. (2008). Posttraumatic stress disorder and social support in female victims of sexual assault: The impact of spousal involvement on the efficacy of cognitive-behavioral therapy. *Behavior Modification*, *32*, 876-896.

Bonnet, L. K. (2012). *The effects of point-of-view video modeling on symbolic play actions and play-associated language utterances in preschoolers with autism*. Unpublished doctoral dissertation, George Mason University, Fairfax, VA.

Busk, P. L., & Serlin, R. C. (2005), Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Donaldson, J. M., Trahan, M. A., & Kahng, S. W. (2014). An evaluation of procedures to increase cooperation related to hoarding in an older adult with dementia. *Journal of Applied Behavior Analysis*, *47*, 410-414.

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

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

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

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

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153-183). Washington, DC: American Psychological Association.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014, June). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/a0037038.

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

Ferron, J. & Ware, W. (1994). Using randomization tests with responsive single-case designs. *Behavior Research and Therapy*, *32*, 787-791.

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

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes*, *54*, 137-154.

Gafurov, B. S., & Levin, J.R. (2014, Apr.). *ExPRT* (*Excel7 Package of Randomization Tests*): *Statistical Analyses of Single-Case Intervention Data* (Version 1.2). Downloadable from https://code.google.com/p/exprt/.

Hanley, G. P., Jin, C. S., Vanselow, N. R., & Hanratty, L. A. (2014). Producing meaningful improvements in problem behavior of children with autism via synthesized analyses and treatments. *Journal of Applied Behavior Analysis*, *47*, 16-36.

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*. *3*, 51-64.

Holden, G., Bearison, D. J., Rode, D. C., Kapiloff, M. F., Rosenberg, G., & Rosenzweig, J. (2002). The impact of a computer network on pediatric pain and anxiety. *Social Work and Health Care*, *36*, 21-33.

Horner, R. H., & Odom, S.L. (2014). Constructing single-case research Designs: Logic and options. In T. R. Kratochwill and J. R. Levin (Eds.), *Single-*

case intervention research: Methodological and statistical advances (pp. 27-51). Washington, DC: American Psychological Association.

Jones, B. F., & Hall, J. W. (1982). School applications of the mnemonic keyword method as a study strategy by eighth graders. *Journal of Educational Psychology*, *74*, 230–237.

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

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, *3*, 206–217.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M, & Shadish, W. R. (2010). Single-case designs technical documentation. In *What Works Clearinghouse: Procedures and standards handbook (Version 2.0)*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26-38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15, 122-144.

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological considerations. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91-125). Washington, DC: American Psychological Association.

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. (1989). A transfer-appropriate-processing perspective of pictures in prose. In H. Mandl & J. R. Levin (Eds.), *Knowledge acquisition from text and pictures* (pp. 83-100). Amsterdam: North-Holland.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243.

Levin, J. R. (1997). Overcoming feelings of powerlessness in aging researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

Levin, J. R., Evmenova, A. S., & Gafurov, B. S. (2014). The single-case data-analysis *ExPRT* (*Excel Package of Randomization Tests*). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 185-219). Washington, DC: American Psychological Association.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB…AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*, *50*, 599-624.

Levin, J. R., Lall, V. F., & Kratochwill, T. R. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology*, *49*, 55-79.

Levin, J. R., Levin, M. E., Cotton, J. W., Bartholomew, S., Hasty, K., Hughes, C., & Townsend, E. A. (1990). What do college students learn from and about an innovative vocabulary-learning strategy? In S. A. Biggs (Ed.), *Innovative learning strategies*, *1989-1990* (pp. 186-206). Pittsburgh, PA: College Reading Improvement Special Interest Group of the International Reading Association.

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.

Lojkovic, D. (2014, Jan.). *Development and use of a modified texting app to increase instances of independent expressive communication for individuals with moderate to severe intellectual and developmental disabilities.* Paper presented at the annual meeting of the Division on Autism and Intellectual Disabilities (DADD), Clearwater, FL.

Luczynski, K. C., Hanley, G. P., & Rodriguez, N. M. (2014). An evaluation of the generalization and maintenance of functional communication and selfcontrol skills with preschoolers. *Journal of Applied Behavior Analysis*, *47*, 246-263.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*, 301-321.

Manolov, R., Evans, J., Gast, D., & Perdices, M. (Eds.). (2014). Single-case experimental design methodology. Special Issue of *Neuropsychological Rehabilitation*, *24*.

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*, 201-215.

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

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal*, *7*, 397-421.

McAllister, L. W., Stachowiak, J.G., Baer, D. M., & Conderman, L. (1969). The application of operant conditioning techniques in a secondary school classroom. *Journal of Applied Behavior Analysis*, *2*, 277-285.

McCleary, R., & Welsh, W. N. (1992). Philosophical and statistical foundations of time-series experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New developments for psychology and education* (pp. 41-91). Hillsdale, NJ: Erlbaum.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*, 191-211.

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. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy*, *32*, 783-786.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*, 357-367.

Pellecchia, M., Connell, J. E., Eisenhart, D., Kane, M., Schoener, C., Turkel, K., Riley, M, & Mandell, D. S. (2011). We're all in this together now: Group performance feedback to increase classroom team data collection. *Journal of School Psychology*, *49*, 411-431.

Regan, K. S., Mastropieri, M. A., & Scruggs, T. E. (2005). Promoting expressive writing among students with emotional and behavioral disturbance via dialogue journals. *Behavioral Disorders*, *33-50*.

SAS (2013). *SAS/IML® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A *d*-statistic for single-case designs that is equivalent to the usual between-groups *d*-statistic. *Neuropsychological Rehabilitation*, *24*, 529-553.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*, 385-405. DOI: 10.1037/a0032964

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971-980.

Stocks, J. T., & Williams, M. (1995). Evaluation of single subject data using statistical hypothesis tests versus visual inspection of charts with and without celeration lines. *Journal of Social Research*, *20*, 105-126.

Woodfield, T. J. (1988). Simulating stationary Gaussian ARMA time series. *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface*, pp. 612 - 617.