Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 23

5-15-2020

An Improved Two Independent-Samples Randomization Test for Single-Case AB-Type Intervention Designs: A 20-Year Journey

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, J. R., Ferron, J. M., & Gafurov, B. S. (2019). An improved two independent-samples randomization test for single-case AB-type intervention designs: A 20-year journey. Journal of Modern Applied Statistical Methods, 18(1), eP3311. doi: 10.22237/jmasm/1556670480

An Improved Two Independent-Samples Randomization Test for Single-Case AB-Type Intervention Designs: A 20-Year Journey

Cover Page Footnote

We wish to acknowledge the early contributions of Bruce Wampold and Venessa Lall (as reflected respectively in Levin & Wampold, 1999; and Lall & Levin, 2004), which provided the impetus for our development of the present single-case two independent-samples randomization-test procedure.
Correspondence concerning this article should be addressed to Joel R. Levin at jrlevin@u.arizona.edu.

Erratum

A previous version of this article incorrectly gave the month of publication of Gafurov & Levin (2020) as April. It has been corrected to March.

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP3311.
doi: 10.22237/jmasm/1556670480

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

An Improved Two Independent-Samples Randomization Test for Single-Case AB-Type Intervention Designs: A 20-Year Journey

Joel R. Levin
University of Arizona
Tuscon, AZ

John M. Ferron
University of South Florida
Tampa, FL

Boris S. Gafurov
George Mason University
Fairfax, VA

Detailed is a 20-year arduous journey to develop a statistically viable two-phase (AB) single-case two independent-samples randomization test procedure. The test is designed to compare the effectiveness of two different interventions that are randomly assigned to cases. In contrast to the unsatisfactory simulation results produced by an earlier proposed randomization test, the present test consistently exhibited acceptable Type I error control under various design and effect-type configurations, while at the same time possessing adequate power to detect moderately sized intervention-difference effects. Selected issues, applications, and a multiple-baseline extension of the two-sample test are discussed.

Keywords: Single-case intervention research, randomization test, two independent samples

Introduction

In recent years, concerted efforts were made from various perspectives to increase the experimental quality and associated scientific credibility of single-case intervention research. Specifically, from a methodological standpoint, more rigorous design standards have been developed (e.g., Gast & Ledford, 2014; Horner & Spaulding, 2010; Kratochwill et al., 2013; Kratochwill & Levin, 2010; Tate et al., 2016), which are increasingly being accepted by the single-case research community. From a data-analysis standpoint, more sophisticated graphical and statistical procedures (e.g., Dart & Radley, 2017; Ferron & Jones, 2006;

doi: 10.22237/jmasm/1556670480 | Accepted: September 10, 2019; Published: May 15, 2020.
Correspondence: Joel R. Levin, jrlevin@u.arizona.edu

Kratochwill & Levin, 2014; Wolfe et al., 2019; Shadish, 2014) have been appearing in the single-case intervention research literature.

In the present note, we add to the single-case researcher's design-and-analysis toolkit what may be an invaluable statistical procedure. The procedure conforms to the methods of analysis classified as single-case randomization tests (Edgington, 1975; Levin et al., 2014), which have also been gaining visibility and respectability over the past quarter of a century (e.g., Craig & Fisher, 2019; Edgington, 1996; Ferron & Levin, 2014; Heyvaert & Onghena, 2014; Michiels & Onghena, 2018). A randomization test is a probability-based nonparametric approach founded on fewer stringent distributional assumptions than standard parametric methods in certain applications, such as with small sample sizes and/or with autocorrelated (serially dependent) outcome observations (e.g., Ferron & Levin, 2014; Levin, 2007). When properly implemented in a manner consistent with a study's design-randomization process, a randomization test yields statistical conclusions that are probabilistically valid (cf. Edgington, 1996; Levin, Kratochwill, & Ferron, 2019; and as will become apparent throughout this article).

The procedure presented here, a two independent-samples randomization test developed to compare two conditions or interventions in single-case AB designs, was inspired by an earlier failure. This procedure will be shown to be statistically valid, in the sense of its exhibiting firm control of the experimental Type I error probability – in contrast to Levin and Wampold's (1999) original version of such a test, which generally did not control the one-tailed Type I error probability to an acceptable degree and thereby produced illusory statistical power results (Lall & Levin, 2014). Then, with its statistical validity intact, the new procedure's practical utility will be examined, in terms of its realistic ability to detect between-samples A-B phase differences of varying types and magnitudes (i.e., its statistical power to detect different varieties of group-by-phase "interaction" effects).

Single-Case AB-Type Intervention Designs

Before continuing, consider single-case AB intervention designs (Levin et al., 2014). They are in the class of "interrupted time-series designs" (e.g., Glass et al., 1975, p. 2), where there are A and B phases, each consisting of multiple outcome observations (O_1, O_2, …, O_P). The A phase typically represents a baseline or control phase and the B phase typically represents an intervention phase, although A and B could also represent two different intervention conditions. A participant (or case) goes through both phases and change is assessed by comparing the set of B observations with the set of A observations with respect to some summary measure

of interest (e.g., within-phase mean, slope, variability). In this sense, when a phase difference in means is the focus, the design is the single-case analog of a conventional one-sample pretest-posttest design. Elevating an AB-type design's acceptability requires the addition of more A and B phases (e.g., ABAB…AB) and/or the addition of more cases (Kratochwill et al., 2013), along with the incorporation of various forms of design-and-analysis randomization (Ferron & Levin, 2014; Kratochwill & Levin, 2010).

For a two-phase AB design, randomization of the A and B phases for each case would appear to be a minimum design requisite, to disentangle the obvious confounding of the intervention and potential time-related effects (e.g., order, testing, fatigue, maturation, history, and the like – see, for example, Shadish et al., 2002). In Levin et al. (2014), concern for this issue was effectively controlled for through a design in which both the A and B phases and the intervention "start points" (i.e., the points of transition from Phase A to Phase B) were randomized on a case-by-case basis, while at the same time capitalizing on the randomization process to increase substantially the power of the study's statistical analysis.

The Levin-Wampold Independent- and Paired-Case Two-Intervention Randomization-Test Models

Levin and Wampold (1999) developed single-case AB two-intervention randomization design-and-analysis models – essentially independent-case and paired-case – which are respectively akin to conventional split-plot and randomized blocks ANOVA designs and analyses (e.g., Kirk, 1995). In the independent-case variation, from a total of N cases, N_X and N_Y cases are randomly assigned to two intervention (or to intervention and control) conditions, X and Y, where for each condition there is a within-case baseline (A) and intervention (B) phase. The design can also be regarded as the single-case analog of a conventional two-group pretest-posttest design (Shadish et al., 2002). In the paired-case variation, the N cases are randomly assigned in pairs to either Intervention X or Intervention Y and all cases go through the A and B phases. In both design variations, a randomization test (the "comparative intervention effectiveness" test) to assess the intervention type (X, Y) by phase (A, B) interaction reveals the critical effect of interest: namely, the differential impact of the two interventions. Because of the specific randomization components implemented in Levin and Wampold's two test variations, it can be argued that each affords a scientifically credible single-case assessment of the comparative effectiveness of two alternative interventions, or of an intervention and a control condition (see, for example., Levin, 1994). (Levin and Wampold's two

models also provide randomization tests of the "general intervention effectiveness" of the two interventions, as defined by the A- to B-phase change in the outcome measure averaged across the two intervention types. The general intervention effectiveness test was not a focus of the present investigation.)

A 20-Year Journey

Unfortunately, often what appears lustrous in theory may lack luster in application. Lall and Levin's (2004) Monte Carlo simulations corroborated that although Levin and Wampold's (1999) paired-case comparative intervention effectiveness test consistently performed exactly as was expected with respect to its one-tailed Type I error (α) control, as was noted earlier, the independent-case comparative intervention effectiveness test did not behave well under most simulation conditions, as represented by combinations of series length, number of cases, number of potential intervention start points, and degree of within-phase "autocorrelation" (Lall & Levin, 2004). With nominal αs set at .05, the test sometimes produced empirical αs as high as .15. As a result, the Levin-Wampold independent-case comparative intervention effectiveness test lacks statistical conclusion validity (Shadish et al., 2002) and therefore cannot be endorsed for widespread practical application.

Faulty first principles Statistically valid randomization tests require a direct correspondence between the random-assignment process and the distribution of all possible randomization outcomes produced by that process (Edgington, 1980; Edgington & Onghena, 2007; Ferron & Levin, 2014; Levin, Ferron, & Gafurov, 2019; Michiels & Onghena, 2018). Without that correspondence, the statistical properties of the test can be seriously compromised – and, of present concern, the ability of the test to control its αs at the desired level (see, for example, Ferron et al., 2003), Levin and Wampold (1999) faced the daunting task of deriving an appropriate randomization distribution in their development of an independent-case comparative intervention effectiveness test. Because there was no readily applicable recipe of how to produce that test's randomization distribution, those researchers worked on the overall objective essentially from a combination of logical inference and brute force. In so doing, they constructed a null randomization distribution for comparing the mean Phase A-to-Phase B change for the two interventions, X and Y, by admitting into the distribution certain logically consistent outcomes while censoring and excluding logically inconsistent ones that were declared inadmissible (cf. Appendix A of Levin & Wampold, 1999).

Unfortunately, that approach proved to be unsuccessful, as was documented by the previously mentioned comparative intervention effectiveness test's unacceptable α-inflated simulation results reported by Lall and Levin (2004).

Sixteen years of frustration and fidgeting Over the past 16 years, two recurring questions effected a good deal of torment by the test and fiddling with the test's randomization distribution:

1. Why didn't Levin and Wampold's (1999) original comparative intervention effectiveness test's null randomization distribution pan out, as had repeatedly been experienced with earlier developed single-case randomization tests of the same class?
2. Are there modifications of the test that can be made to produce an appropriate randomization distribution and enable the test to function properly?

Finally A 2019 "Aha!" moment occurred when rather than conducting microanalyses of admissible and inadmissible randomization-distribution outcomes, our attention targeted the formulation of a back-to-the-drawing-board conceptual model on which the single-case comparative intervention effectiveness test was based. A systematic analysis of the problem revealed that an incorrect model had been applied to generating the proper randomization distribution for this test. To right the wrong, it was necessary to reconstruct the Levin-Wampold comparative intervention effectiveness randomization test from a different perspective. That perspective arose from considering the basis of two-sample permutation and randomization tests in the traditional nonparametric statistical literature (e.g., Conover, 1999; Levin, 2007).

Step 1. Specifically, if N participants are to be randomly assigned to two intervention (or to intervention and control) conditions (X and Y), with N_X participants in one condition and N_Y participants in the other, there is a total of $N! / (N_X! \, N_Y!)$ possible assignments of participants to conditions. For an example based on $N = 6$ participants and $N_X = N_Y = 3$ then, $6! / (3! \, 3!) = 20$ different assignments of 3 participants to each condition are possible. The same assignment process is applied in the present single-case context and a test statistic defined as the difference between the B- and A-phase means averaged across the three cases in one condition for each of the 20 possible combinations can be calculated. So, for example, the 20 Condition X B-A mean differences would consist of those associated with: Cases 1, 2, and 3; Cases 1, 2, and 4; etc., all the way through Cases

4, 5, and 6. 2. As a relevant aside, for the comparative intervention effectiveness tests discussed here, the B-A mean-difference outcomes from only one of the two conditions needs to be considered because the other condition's outcomes are the mirror-image complements.

Step 2. With this replicated single-case design that adopts a randomized intervention start-point rationale, we also take into account the number of potential intervention start points for each case (Marascuilo & Busk, 1988). With k_i such start points for each of the N cases, the number of total randomization-distribution outcomes produced is equal to

$$k_1 \times k_2 \times \mathrm{K} \times k_N = \prod_{i=1}^{N} k_i \ ,$$

which for equal k_i reduces to k^N (see Levin et al., 2014). In this example, a total of $N = 6$ total participants, each of whom is randomly assigned $k = 2$ potential intervention start points, would yield $2^6 = 64$ distinct randomization outcomes.

Step 3. The total number of outcomes in the Condition X randomization distribution then becomes the product of the results in Step 1 and Step 2. Applying this multiplication operation produces a total of

$$\frac{N!}{N_{\mathrm{X}}!N_{\mathrm{Y}}!} \times \prod_{i=1}^{N} k_i$$

randomization-distribution outcomes, which, for the present example is equal to $20 \times 64 = 1280$ (i.e., each of the 20 3-case Condition X groupings combined with the 2 potential intervention start-point possibilities for each case). Accordingly, with a Type I error of .05, if the actually obtained B-A mean difference in the randomization distribution is among the $.05 \times 1280 = 64$ largest in the predicted direction (e.g., X > Y), then a one-tailed significance probability (or p-value) of $p \leq .05$ can be claimed.

An Assessment of the New Test's Statistical Behavior

With the development of the new single-case comparative intervention effectiveness randomization test ostensibly on firm ground, we conducted a Monte Carlo simulation study to provide empirical support for its statistical-conclusion validity with respect to Type I error and power.

Method

Incorporating the three steps just described, Monte Carlo simulation methods were implemented to examine the new comparative intervention effectiveness test's Type I error rates and power for the test for: (*a*) designs consisting of 6 cases, of which 3 were randomly assigned to each of the two groups; as well as for (*b*) designs consisting 8 cases, of which 4 were randomly assigned to each of the two groups. We varied the series lengths from 10 to 30, which cover the range of series lengths that have been included in most single-case intervention studies (e.g., Ferron et al., 2010). In designs with 6 cases and 10 observations per case, we examined conditions where the actual intervention start point was established according to three different configurations: (1) it was set at Observation 6 for all cases (i.e., the same single fixed intervention start point was used for each case); (2) it was randomly selected from the same two potential intervention start points for each case (namely, Observation 5 or 6); or (3) it was randomly selected from the same five potential intervention start points for each case (namely, Observation 4, 5, 6, 7, or 8).

Similarly, for designs with 6 cases and 30 observations per case, conditions were examined where the intervention actual intervention start point was set at Observation 16 for all cases, with the preceding configurations being: (1) Observation 16 for each case; (2) Observation 15 or 16 for each case; or (3) Observation 14, 15, 16, 17, or 18 for each case. For conditions with a single fixed intervention start point, the number of possible random assignments was 20 [i.e., 6! / (3! 3!)]; for conditions with 2 potential intervention start points, the number of possible random assignments was 1,280 (20×2^6); and for conditions with 5 potential intervention start points, the number of possible random assignments was 312,500 (20×5^6). For designs with 8 cases, parallel conditions were set for designs with 1 or 2 potential intervention start points (yielding 70 and 17,920 possible intervention start points, respectively), but we did not examine designs with 5 possible intervention start points (yielding 27,343,750 possible random assignments), because of the excessive amount of computing space and time required.

For each of these designs, time series data for each case were generated by adding a series of errors (e) to a series of true values (μ) such that at time *t* for case *i* the outcome value was $y_{ti} = \mu_{ti} + e_{ti}$. The errors were generated for each time series using the autoregressive moving-average simulation function (ARMASIM) in SAS. A first-order autoregressive model $e_t = \rho e_{t-1} + a_t$ was specified where the variance of the white noise, VAR(a_t), was set to 1.0 and the autocorrelation, ρ, was set to .00

or .30. These values were used in other simulations of multiple-baseline data (e.g., Ferron & Sentovich, 2002; Ferron & Ware, 1995; Levin et al., 2018) and range from no autocorrelation (i.e., $\rho = 0$) to a value that is a little larger than the single-case intervention research literature-based, meta-analytic average bias-adjusted autocorrelation of .20 (Shadish & Sullivan, 2011).

The true values were based on both stable baseline phases ($\mu_{ti} = 0$) and stable intervention phases ($\mu_{ti} = d_i$). Thus, for all $d_i > 0$ (non-null B-phase) conditions, immediate abrupt effects that remained constant throughout the intervention phase were assumed. The values of d_i for the first intervention condition, X (d_X) and for the second intervention condition, Y (d_Y) were set to values to obtain four effect-size combinations: (1) null ($d_X = 0$, $d_Y = 0$); (2) consistent small effect ($d_X = 1$, $d_Y = 1$); (3) consistent large effect ($d_X = 3$, $d_Y = 3$); and (4) six differential effect sizes favoring Condition X (described in the Results condition). For the new comparative intervention effectiveness test, the null and the two consistent effect conditions permitted estimating the new test's Type I error rate (α), whereas the differential effect condition permitted examining the new test's power.

For configurations with 1 or 2 intervention start points, 100,000 experiments were simulated for each condition. For configurations with 5 intervention start points, 10,000 experiments were simulated per condition. For the latter configurations, we sacrificed a little precision to accommodate the substantial increase in the number of permutations and processing time required for simulated experiments based on 5 potential intervention start points per case.

Results

All statistical tests conducted were directional (one-tailed) based on $\alpha = .05$. One-tailed tests have typically been adopted for single-case intervention simulation research (e.g., Ferron & Levin, 2014; Levin et al., 2018) for both applied and related statistical reasons. From an applied perspective, single-case interventionists generally have – and should have – well-articulated knowledge about the two experimental conditions that they wish to compare, be they an intervention and a nonintervention control condition or two different intervention conditions (e.g., Horner & Spaulding, 2010; Kazdin, 2011; Kratochwill et al., 2013). Consequently, single-case interventionists are positioned to make strong better than or worse than predictions regarding their anticipated intervention outcomes. Relatedly, because single-case research, by definition, is characterized by small sample sizes and, relative to conventional group experimental research is generally inferior with respect to statistical power, single-case interventionists are well-advised to conduct

directional tests so as to improve their chances of uncovering intervention effects that might otherwise have gone undetected (at least from an inferential statistical standpoint).

The results for the present two independent-samples comparative intervention effectiveness test are summarized in Tables 1 and 2. These results are straightforward and clearly supportive of the new test's statistical viability, as will now be described.

Type I Error

The first three configurations of Table 1, for which $d_X = d_Y$, present outcomes reflecting empirical αs. Those obtained values consistently reveal strict control of the new test's αs at or below the specified nominal α of .05, with greater conservativeness for the $d_X = d_Y = 3$ configuration than for the two others. Just to be on the safe side, we also examined a selected set of nondirectional (two-tailed) test comparisons based on 6 cases and two potential intervention start points. [With 6 cases and only one potential intervention start point per case, i.e., a total of 20 possible randomization outcomes, the smallest two-tailed empirical α (or p-value) obtainable is 2/20 = .10.] The results duplicated those of the one-tailed test comparisons, maintaining Type I error control (all empirical αs $\leq .05$) and becoming more conservative when $d_X = d_Y = 3$.

Table 1. Type I error rates (based on $\alpha = .05$, one-tailed) for the "comparative intervention effectiveness" test with cases randomly assigned to the two intervention conditions, X and Y

				6 Cases			8 Cases	
d_X	d_Y	ρ	SL	1 SP	2 SP	5 SP	1 SP	2 SP
0	0	0.00	10	0.050	0.051	0.051	0.043	0.049
			30	0.050	0.051	0.053	0.043	0.051
		0.30	10	0.050	0.050	0.051	0.043	0.051
			30	0.050	0.049	0.052	0.042	0.051
1	1	0.00	10	0.049	0.047	0.044	0.042	0.048
			30	0.050	0.049	0.048	0.043	0.050
		0.30	10	0.049	0.047	0.045	0.042	0.048
			30	0.049	0.050	0.050	0.044	0.049
3	3	0.00	10	0.049	0.028	0.013	0.042	0.032
			30	0.050	0.038	0.016	0.041	0.041
		0.30	10	0.051	0.034	0.020	0.042	0.038
			30	0.049	0.042	0.029	0.043	0.046

Note: d_X = effect size for group X, d_Y = effect size for group Y; ρ = autocorrelation; SL = series length; 1, 2, and 5 SP = 1, 2, and 5 potential intervention start points for each case, respectively

Table 2. Power (based on α = .05, one-tailed) for the "comparative intervention effectiveness" test with an autocorrelation of ρ = .30 and cases randomly assigned to the two intervention conditions, X and Y

			6 Cases			8 Cases	
d_X	d_Y	SL	1 SP	2 SP	5 SP	1 SP	2 SP
1	0	10	0.339	0.370	0.385	0.419	0.476
		30	0.619	0.641	0.667	0.754	0.802
2	0	10	0.774	0.823	0.844	0.897	0.932
		30	0.981	0.986	0.991	0.998	0.999
3	0	10	0.968	0.983	0.989	0.996	0.999
		30	1.000	1.000	1.000	1.000	1.000
2	1	10	0.341	0.356	0.355	0.418	0.464
		30	0.619	0.641	0.648	0.756	0.801
3	1	10	0.777	0.811	0.813	0.896	0.927
		30	0.981	0.986	0.990	0.998	1.000
3	2	10	0.341	0.332	0.288	0.418	0.442
		30	0.618	0.628	0.628	0.753	0.796

Note: d_X = effect size for group X, d_Y = effect size for group Y; SL = series length; 1, 2, and 5 SP = 1, 2, and 5 potential intervention start points for each case, respectively

Power

Group X was set to benefit more from the introduction of its intervention condition than was Group Y (i.e., from its mean B-A increase), and so all powers are associated with various X > Y effect sizes. Selected results, based on 6 cases equally divided between the two intervention conditions, 1, 2, or 5 potential intervention start points per case, and an autocorrelation of ρ = .30, are presented in Table 2. [These results are not directly comparable to those of Lall and Levin (2004) for Levin and Wampold's (1999) comparative intervention effectiveness test because the present results are based exclusively on immediate abrupt effect types, whereas Lall and Levin's results were averaged across four different effect types (*viz*., immediate abrupt, delayed abrupt, immediate gradual, and delayed gradual).]

As noted in Table 2 for both the 6- and 8-case situations, with a series length of 10 observations, the powers for comparative (differential) effect sizes of 1 (i.e., $d_X - d_Y = 1 - 0$, $2 - 1$, and $3 - 2$) are in the .30s and .40s, and therefore inadequate, with the lowest power value of .29 occurring in the 6-case situation for the $d_X - d_Y = 3 - 2$ effect-size difference with 5 potential intervention start points. However: (1) with a series length of 30 observations the obtained powers in the .60s for the 6-case situation and in the .80s and .90s for the 8-case situation become much more reasonable; and (2) for comparative effect sizes of 2 or 3

($d_X - d_Y = 2 - 0$, $3 - 1$, and $3 - 0$) – which are not uncommon in single-case intervention research (e.g., Ferron et al., 2014; Marquis et al., 2000; Rogers & Graham, 2008) – even the 6-case situation powers are in the high .70s, .80s, and .90s. (The ds ≥ 1 are representative only of d effect sizes in the *published* single-single case literature.) The powers in Table 2 were generated assuming a fairly conservative (from a power perspective) autocorrelation of $\rho = 30$. When the autocorrelation decreases toward 0, all powers in Table 2 increase in a pattern comparable to that for $\rho = .30$. As a few examples based on the 6-case situation and 2 potential intervention start points per case: (1) with a series length of 10, the comparative $\rho = 0$ and $\rho = .30$ powers for effect-size differences of $1 - 0$, $2 - 1$, and $3 - 2$, respectively, are .50 vs. .37, .47 vs. .36, and .44 vs. .33; and (2) with a series length of 30 and the same effect-size differences, they are .85 vs. .64, .85 vs. .64, and .84 vs. .63.

General Intervention Effectiveness Test

The comparative intervention effectiveness test corresponds to a treatment-by-time interaction in conventional group design research. Often of additional interest in the group design context is the time main effect: that is, whether there is a mean change from Time 1 (e.g., pretest) to Time 2 (posttest) averaged across the two treatment conditions. Such a test in the present single-case design context is available through either the Levin-Wampold (1999) general intervention effectiveness test or the equivalent Marascuilo-Busk (1988) replicated AB design procedure. Specifically, the test assesses whether there is a change in outcomes (here, a change in levels, or means) from the A-to-B phase across the N cases in the study (i.e., ignoring the X or Y experimental condition to which the cases were assigned). The test yields a total of

$$k_1 \times k_2 \times \mathrm{K} \times k_N = \prod_{i=1}^{N} k_i$$

(or k^N for equal k_i) randomization distribution outcomes as reflected in Step 2 of the present comparative intervention effectiveness procedure) – where, again, k_i represents the number of potential intervention start points for the i^{th} case. The general intervention effectiveness test has previously been found both to maintain strict Type I error control and to produce acceptable powers for detecting effects of typical interest to single-case intervention researchers (e.g., Ferron & Sentovich, 2002; Lall & Levin, 2004; Levin et al., 2014).

Discussion

It is apparent that our 20-year two independent-samples comparative intervention effectiveness randomization-test journey ended not with a whimper but a bang. In contrast to the earlier version of the test, the present version consistently maintained acceptable Type I error control, while exhibiting adequate power with a total of 6 cases – and especially with 8 cases – equally divided between two intervention conditions, to detect a variety of between-samples intervention effects of moderate size (i.e., $d_X - d_Y \geq 2$) under reasonably realistic outcome-autocorrelation values of .30. Somewhat unexpectedly, and as may appreciated from the Table 2 results, the most dramatic power-enhancing factor proved to be the length of the series: specifically, as the series length increased from 10 to 30 outcome observations. At the same time, certain uncharted territories for the present comparative intervention effectiveness randomization test have yet to be fully explored.

Adaptation to Two Independent-Samples Multiple-Baseline Designs

Likely among the most appealing to single-case intervention researchers would be adapting the present AB-design procedure for application in multiple-baseline designs (see, for example, Levin et al., 2018). Encouragingly, the approach reported here can be directly imported to a multiple-baseline design by instead of defining a common range based on k potential intervention start points for each case, the k start points for each case would be systematically staggered in multiple-baseline fashion, thereby yielding the same number of possible randomization-test outcomes, namely, $N! / (N_X!\, N_Y!) \times k^N$. With k_i potential intervention start points for each case, this becomes

$$\frac{N!}{N_X!N_Y!} \times \prod_{i=1}^{N} k_i \, .$$

However, the resulting test will provide only a partially complete two-sample multiple-baseline analysis because the random assignment of cases to the N_X and N_Y levels (or tiers) of the two intervention conditions would not be taken into account in the analysis – in contrast to how it is cleverly effected in the multiple-baseline tier-permutation approach developed by Wampold and Worsham (1986). Implementing such "case randomization" (Ferron & Levin, 2014) is an order of magnitude more challenging than in the present AB design because it requires, for each intervention condition, a consideration of both: (1) the cases' stagger positions

(1 to N_p) and (2) the number of potential intervention start points for each case (1 to N_k), the latter of which increases the number of possible randomization outcomes by a factor of ($N_X! \times N_Y!$), bringing the total number of randomization-distribution outcomes to $N! \times k^{N,}$ or, in general,

$$N! \times \prod_{i=1}^{N} k_i$$

(Although the multiple-baseline extension of the two-sample procedure is beyond the purview of the current investigation, an assessment of its statistical properties is currently underway by the present authors.)

Connection to Conventional "Group" Randomization and Permutation Tests

With only one fixed intervention start point for each case, the present test is equivalent to a conventional group two-sample exact randomization test (with random assignment to groups) or permutation test (without random assignment to groups) and for large enough sample sizes, to a parametric two-sample *t* test, when applied to the *N* cases' B-A mean differences – for related examples and discussion, see Ferron & Levin (2014) and Levin (2007). Both tests will be associated with $N! / (N_X! N_Y!)$ randomization distribution outcomes. However, adding *k* potential intervention start points for each case to the present procedure increases the number of randomization outcomes by a factor of k^N, and generally (though not invariably) with it the associated statistical power, as is evidenced by the results in Table 2.

Consideration of the Present Effect Types

The present mean/level simulations were conducted assuming a stable baseline (A) phase for both Groups X and Y, followed by either: (a) a continuing stable intervention (B) phase at the same level as baseline for both of the groups; (b) or an immediate abrupt B-phase increase in level in either or both of the groups. With intervention-phase effects that are delayed or gradual, one can expect the powers reported in Table 2 to be lower – and typically, considerably lower (Levin et al., 2018). However, if in advance a single-case intervention researcher can correctly anticipate the nature of these effect types, then specific ameliorative adjusted measures can be constructed to lessen the amount of lost power (Levin et al., 2017). Of the various effect types previously examined, detecting gradual, rather than abrupt, changes in A- to B-phase levels poses the most severe loss-of-power

problems. In addition, a different set of considerations is required if the researcher's focus is on phase-change differences in trend/slope or variability, rather than on differences in level (Levin, Ferron, & Gafurov, 2019).

Caution About Nonrandomly Formed Groups

The present two independent-samples procedure should not be implemented in situations where Groups X and Y are nonrandomly constituted, such as when the two groups consist of cases that represent demographic, classification, or status variables (e.g., age/grade, gender, ability level) and comparisons of the two groups are made on some task or measure (e.g., the comparative effectiveness of an instructional intervention for students with and without a learning disability). In such instances, including the combinatorial randomized group-formation portion of the procedure [*viz.*, $N! / (N_X! \, N_Y!)$] is invalid and would result in an overdetermination of the legitimate number of possible randomization-distribution outcomes. Currently being explored is whether and how our two-group test can be adapted for legitimate application in nonrandom-assignment-to-groups contexts.

Concluding Comments

The present experimental expedition concludes with a few comments. First, to preserve our two independent-samples test's internal validity, cases must be randomly assigned to the study's administration start times, so as not to confound between-condition comparative intervention effectiveness differences with time or order differences associated with the intervention conditions. Second, whether our two-sample procedure is equally well suited for behaviorally based observational designs and cognitively based acquisition designs has yet to be determined. Without going into details here, that is because with the latter design types, the random assignment of intervention start points to cases within the two conditions could end up producing complicated phase-by-content interpretations in the two conditions. Third, and as was noted earlier, single-case randomization tests for the paired-case variation of the two-intervention design have been developed (Levin & Wampold, 1999). So too have randomization tests for AB crossover designs (Levin et al., 2014) and alternating treatment designs, which are both applicable for within-case comparisons of different interventions (Levin et al., 2012).Finally, each of these tests – including, in particular, the new two independent-samples randomization test – can be executed through the freely accessible, downloadable *ExPRT* (*Excel Package of Randomization Tests*) Version 4.1 statistical software (Gafurov & Levin, 2020). All told, we these procedures have the potential to be valuable, scientifically

credible, and statistically sound design-and-analysis strategies for single-case interventionists to consider in their research investigations.

Acknowledgement

The first two authors contributed equally to this study. We wish to acknowledge the early contributions of Bruce Wampold and Venessa Lall (as reflected respectively in Levin & Wampold, 1999; and Lall & Levin, 2004), which provided the impetus for our development of the present single-case two independent-samples randomization-test procedure. Correspondence concerning this article should be addressed to Joel R. Levin at jrlevin@u.arizona.edu.

References

Craig, A. R., & Fisher, W. W. (2019). Randomization tests as alternative analysis methods for behavior analytic data. *Journal of the Experimental Analysis of Behavior*, *111*(2), 309-328. doi: 10.1002/jeab.500

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd edition). New York: Wiley.

Dart, E. H., & Radley, K. C. (2017). The impact of ordinate scaling on the visual analysis of single-case data. *Journal of School Psychology*, *63*, 105-118. doi: 10.1016/j.jsp.2017.03.008

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology*, *90*(1), 57-58. doi: 10.1080/00223980.1975.9923926

Edgington, E. S. (1980). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics*, *5*(3), 261-267. doi: 10.3102/10769986005003261

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, *34*(7), 567-574. doi: 10.1016/0005-7967(96)00012-5

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

Ferron, J., Farmer, J., & Owens, C. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel modeling approaches. *Behavior Research Methods*, *42*, 930-943. doi: 10.3758/brm.42.4.930

Ferron, J., Foster-Johnson, L., & Kromrey, J. D. (2003). The functioning of single-case randomization tests with and without random assignment. *Journal of Experimental Education*, *71*(3), 267-288. doi: 10.1080/00220970309602066

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education*, *75*(1), 66-81. doi: 10.3200/jexe.75.1.66-81

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. doi: 10.1037/14376-006

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating casual effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*(4), 493-510. doi: 10.1037/a0037038

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, *70*(2), 165-178. doi: 10.1080/00220970209599504

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education*, *63*(2), 167-178. doi: 10.1080/00220973.1995.9943820

Gafurov, B. S., & Levin, J. R. (2020, March). *ExPRT - Excel® package of randomization tests: Statistical analyses of single-case intervention data* (Version 4.1). https://ex-prt.weebly.com/

Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology* (2nd edition). New York: Routledge. doi: 10.4324/9780203521892

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments*. Boulder, CO: University of Colorado Press.

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*(1), 51-64. doi: 10.1016/j.jcbs.2013.10.002

Horner, R., & Spaulding, S. (2010). Single-case research designs. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1386-1394). Thousand Oaks, CA: Sage Publications.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd edition). New York: Oxford University Press.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd edition). Pacific Grove, CA: Brooks/Cole.

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*(1), 26-38. doi: 10.1177/0741932512452794

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 122-144. doi: 10.1037/a0017736

Kratochwill, T. R., & Levin, J. R. (Eds.) (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association. doi: 10.1037/14376-000

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*(1), 61-86. doi: 10.1016/j.jsp.2003.11.002

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243. doi: 10.1007/bf02213185

Levin, J. R. (2007). Randomization tests: Statistical tools for assessing the effects of educational interventions when resources are scarce. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 115-123). Greenwich, CT: Information Age.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods*,*13*(2), 2-52. doi: 10.22237/jmasm/1414814460

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology*, *63*, 13-34. doi: 10.1016/j.jsp.2017.02.003

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*, *21*(5), 290-311. doi: 10.1080/17518423.2016.1197708

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2019). *Randomization tests of trend and variability for single-case multiple-baseline designs* (Unpublished manuscript). Tuscon, AZ: University of Arizona.

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*(5), 599-624. doi: 10.1016/j.jsp.2012.05.001

Levin, J. R., Kratochwill, T. R., & Ferron, J. M. (2019). Randomization procedures in single-case intervention research contexts: (Some of) "the rest of the story". *Journal of Applied Behavior Analysis, 112*(3), 334-348. doi: 10.1002/jeab.558

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly*, *14*(1), 59-93. doi: 10.1037/h0088998

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

Marquis, J. G., Horner, R. H., Carr, E. G., Turnbull, A.P., Thompson, M., Behrens, G. A., Magito-McLaughlin, D., McAtee, M. L., Smith, C. E., Ryan, K. A., & Doolabh, A. (2000). A meta-analysis of positive behavior support. In R. Gersten, E. P. Schiller, & S. Vaughn (Eds.), *Contemporary special education research: Syntheses of knowledge base on critical instructional issues* (pp. 137-178). Mahwah, NJ: Erlbaum.

Michiels, B., & Onghena, P. (2018). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*, 2454-2476. doi: 10.3758/s13428-018-1084-x.

Rogers, L. A., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, *100*(4), 879-906. doi: 10.1037/0022-0663.100.4.879

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*(2), 109-122. doi: 10.1016/j.jsp.2013.11.009

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

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. doi: 10.3758/s13428-011-0111-y

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., … Wilson, B. (2016). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016 statement. *Evidence-Based Communication Assessment and Intervention*, *10*(1), 44-58. doi: 10.1080/17489539.2016.1190525

Wampold, B., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, *8*(2), 135-143.

Wolfe, K., Dickenson, T. S., Miller, B., & McGrath, K. V. (2019). Comparing visual and statistical analysis of multiple baseline design graphs. *Behavior Modification*, *43*(3), 361-388. doi: 10.1177/0145445518768723