Additional Comparisons of Randomization-Test Procedures for Single-Case Multiple-Baseline Designs: Alternative Effect Types

Joel R. Levin

University of Arizona

John M. Ferron

University of South Florida

Boris S. Gafurov

George Mason University

Author Note

The first two authors contributed equally to this study. We are grateful to the reviewers of earlier versions of this article for their helpful revision suggestions. Correspondence concerning the article should be addressed to Joel R. Levin at jrlevin@u.arizona.edu.

Abstract

A number of randomization statistical procedures have been developed to analyze the results from single-case multiple-baseline intervention investigations. In a previous simulation study, comparisons of the various procedures revealed distinct differences among them in their ability to detect immediate abrupt intervention effects of moderate size, with some procedures (typically those with randomized intervention start points) exhibiting power that was both respectable and superior to other procedures (typically those with single fixed intervention start points). In Investigation 1 of the present follow-up simulation study, we found that when the same randomization-test procedures were applied to either delayed abrupt or immediate gradual intervention effects: (1) the powers of all of the procedures were severely diminished; and (2) in contrast to the previous study's results, the single fixed intervention start-point procedures generally outperformed those with randomized intervention start points. In Investigation 2 we additionally demonstrated that if researchers are able to successfully anticipate the specific alternative effect types, it is possible for them to formulate adjusted versions of the original randomization-test procedures that can recapture substantial proportions of the lost powers.

Additional Comparisons of Randomization-Test Procedures for Single-Case Multiple-Baseline Designs: Alternative Effect Types

Over the past several years single-case research methodology and associated data-analysis procedures have elevated their scientific "credibility" (Levin, 1994) among educational and psychological intervention researchers (see, for example, Kratochwill et al., 2013; and Kratochwill & Levin, 2014). The once-common single-case two- and three-phase AB and ABA designs, respectively, are now considered to be lacking scientific validity (Kratochwill et al., 2013) and so designs such as the ABAB "reversal" design, the alternate treatments design, and the multiple-baseline design have been advocated in their stead. Of these, many single-case intervention researchers (including the present authors) believe that the systematically staggered multiple-baseline design possesses the strongest internal-validity characteristics of all commonly adopted single-case designs in terms of its ability to document causal relationships between interventions and outcomes (see, for example, Horner & Odom, 2014; and Levin, 1992).

Yet, despite the high methodological marks accorded to single-case designs such as the multiple baseline, even greater scientific credibility can be attained through an interventionist's implementation of various forms of design randomization and data-analysis randomization to enhance the research's internal validity and statistical conclusion validity, respectively (Kratochwill & Levin, 2010). To emphasize the enhanced scientific credibility that accrues to single-case designs through researcher-managed *randomization* and *control*, and consistent with Shadish, Cook, and Campbell's (2002) orientation, we have started referring to designs that possess these two methodological components as "experimental" single-case intervention research designs (see also de Jong et al., 2008). It is implicit that all of the single-case intervention designs discussed in this article encompass randomization and control as two defining "experimental research" requisites.

The general topic of concern in the present simulation study is on statistical tests applied to data from multiple-baseline intervention studies, with a specific purpose of extending the results of a recently reported study on the statistical properties of a number of multiple-baseline randomization tests (Levin, Ferron, & Gafurov, 2016). The statistical properties in question are Type I error and power, and the randomization tests are ones that have appeared in the single-case literature throughout the past nearly half century. Single-case randomization tests have been gaining traction among educational intervention researchers, and notably, among researchers who have focused on academic and behavioral concerns (e.g., Ainsworth, Evmenova, Behrmann, & Jerome, 2016; Bice-Urbach & Kratochwill, 2016; Bardon, Dona, & Symons, 2008; Brewer & White, 1994; Hwang, Levin, & Johnson, 2016; Lojkovic, 2014; Markham, Porter, & Ball, 2011; Regan, Mastropieri, & Scruggs, 2005). The results of the present simulation study should be of direct relevance to single-case intervention researchers with those concerns.

Different Types of Single-Case Design-and-Analysis Randomization

Single-case intervention designs are interrupted time-series designs (e.g., Glass, Willson, & Gottman, 1975; McCleary, McDowall, & Bartos, in press) with at least two phases, a baseline or control phase (A) and an intervention or experimental phase (B), with each phase generally containing multiple outcome observations (Horner & Odom, 2014). In such designs, cases consist of either individual participants or clustered aggregates such as dyads, small groups, classrooms, communities, etc.; and the successive observations produced by each case typically are not independent, with the degree of nonindependence reflected by the magnitude of the autocorrelation coefficient. Because of the autocorrelated nature of the data, standard statistical procedures for assessing between-phase changes (for example, through parametric *t* or *F* tests) (1) do not satisfy the procedures' requisite assumptions, (2) will generally lead to unwarranted statistical conclusions, and therefore (3) should not be applied (see, for example, Ferron & Levin, 2014). Methodologically stronger single-case intervention

designs are produced when they are replicated across cases—and consequently, commonly applied designs (e.g., AB, reversal, alternating treatment, multiple-baseline) must include a prescribed multiple-case replication component before they are endorsed by single-case intervention research "standards" committees (e.g., Kratochwill et al., 2013).

As we have indicated previously (Levin, Ferron, & Gafurov, 2014), there are currently four different randomization variations that are being incorporated into single-case designs and statistical analyses: within-case intervention/phase-order randomization, between-case intervention randomization, case randomization, and intervention start-point randomization. Within-case intervention/phase-order randomization can be applied in ABAB…AB, alternating treatment, and simultaneous treatment designs (Levin, Ferron, & Kratochwill, 2012); and between-case intervention randomization is manifested in situations where there is random assignment of one or more cases to one particular intervention (X) and another case or cases to a different intervention (Y), as in Levin and Wampold's (1999) "randomized pairs" AB design.

With the present study's focus on multiple-baseline designs, the third and fourth randomization types (case randomization and intervention start-point randomization) comprise the relevant types under consideration here. With case randomization, the *N* cases are randomly assigned to the *N* staggered positions, or "tiers" (e.g., Barton & Reichow, 2012), of the multiple-baseline design. With intervention start-point randomization, the within-position observation associated with the first session of the intervention phase is randomly selected from a pre-established number of potential intervention start points (*k*) that are "acceptable" to the researcher – an innovative process initially proposed by Edgington (1975) that has beneficial randomization-test consequences (Ferron & Levin, 2014). Thus, for example, in a four-case, 25-observation (O) multiple-baseline study, the researcher could specify that a minimum of 5 baseline (A phase) and 5 intervention (B phase) observations are required for each case, with two possible design variations represented by those specifications as follows: (*a*) the case within each multiple-baseline position receives a randomly selected intervention start point

somewhere between O_6 and O_{21} inclusive (Marascuilo & Busk, 1988), resulting in a total of 16 potential intervention start points for each case;[1] or (*b*) the cases that are randomly assigned to Positions 1-4 are randomly assigned intervention start points falling within the intervals O_6-O_9, O_{10}-O_{13}, O_{14}-O_{17}, and O_{18}-O_{21}, respectively (Koehler & Levin, 1998).

Overview of Five Multiple-Baseline Randomization Tests

Five different single-case multiple-baseline randomization-test design-and-analysis procedures (two of which included two different variations) were examined in the present simulation study, with three of the procedures involving intervention start-point randomization: Levin et al.'s (2016) restricted Marascuilo-Busk (1988) procedure, Koehler and Levin's (1998) "regulated randomization" procedure (two variations), and Levin et al.'s modified Revusky (1967) procedure (two variations). The two procedures with a single fixed intervention start point were the Wampold-Worsham (1986) procedure, the predecessor and special-case version of the Koehler-Levin procedure; and the special-case version of the modified Revusky procedure.

An important distinction among the five procedures is whether the procedure is based primarily on what might be referred to as "within-case" comparisons or primarily on "between-case" comparisons (Levin, Evmenova, & Gafurov, 2014; Levin et al., 2016). Within-case comparisons focus on each participant's B- to A-phase "change" on some predesignated outcome of interest (as defined, for example, in terms of "level"/mean, "trend"/slope, or variability—see Horner & Odom, 2014), with those change measures then aggregated across cases to become part of a randomization distribution. Between-case comparisons also focus on the same B- to A-phase across-cases mean changes, but the randomization distribution involves comparing the *N* cases' mean changes at various positions and stages of the multiple-baseline design. Each procedure will now be briefly described in terms of this within- vs. between-case distinction, along with the type(s) of randomization that the procedure incorporates. In the following discussion we will describe the test procedures for the situation in which an abrupt change in level is anticipated and thus the B-A phase mean differences (i.e.,

Phase B minus Phase A) lead to an appropriate comparison of the B and A phases. For the interested reader, a complete description of each procedure's computational underpinnings and worked examples may be found in its original source (see also Levin, Evmenova, & Gafurov, 2014; and Levin et al., 2016); and for convenient application, all of the procedures are freely available for downloading at Gafurov and Levin's (2016) single-case *ExPRT* (*Excel* Package of Randomization Tests) website.

Within-Case Randomization-Test Procedures

Wampold and Worsham (1986) procedure. With this procedure, based on case randomization, the B-A mean difference associated with each case's position in the multiple-baseline design is calculated and then aggregated across cases to define the *actually obtained* overall mean difference. That mean difference is included in a randomization distribution that contains a total of *N*! mean differences, with those differences calculated from the *N*! possible assignments of cases to the *N* multiple-baseline positions. So, for example, with 4 cases, there would be 4! = 24 possible assignments of cases to positions. For simplicity, suppose that the actual assignment was Case A in Position 1, Case B in Position 2, Case C in Position 3, and Case D in Position 4 ($A_1B_2C_3D_4$). Another possible assignment would have been $A_1B_2C_4D_3$ or $A_3B_1C_4D_2$, and so on up to a total of 24 different permutations. For each of those possible assignments, the mean B-A difference is calculated for the point of intervention associated with the specific position number (1, 2, 3, or 4) and then aggregated across cases, resulting in a randomization distribution containing 24 mean differences. If the actually obtained mean difference happened to be the largest of all 24 mean differences, the significance probability for that particular outcome (occurring by chance) is equal to 1/24 = .042, which is less than a standard one-tailed significance level (α) of .05 and would therefore be regarded as statistically significant. If, on the other hand, the actually obtained mean difference were the second largest of all 24 mean differences, the significance probability for obtaining an outcome that large or larger is equal to 2/24 = .083, which is more than .05 (one-tailed) and would therefore be

regarded as statistically nonsignificant. In that regard, we note that all statistical outcomes reported in the present study are derived from one-tailed statistical tests, under the reasonable assumption that single-case interventionists have (or at least should have) firm *a priori* knowledge or predictions about the direction of their anticipated intervention effects.

Restricted Marascuilo-Busk procedure (Levin et al., 2016). Marascuilo and Busk (1988) extended Edgington's (1975) random intervention start-point model for a single case to accommodate multiple independent cases. With the Marascuilo-Busk procedure, a common range of potential intervention start points is designated and randomly sampled from for all N cases.[2] With N cases and k potential intervention start points for each case, the total number of randomization-distribution outcomes is equal to k^N; so, for example, with $N = 4$ cases and k = an 8-observation interval of potential intervention start points for all cases, there would be $8^4 = 4{,}096$ randomization-distribution outcomes.

Even though Marascuilo and Busk (1988) promote their replicated AB-design procedure as also being applicable to multiple-baseline designs, it cannot strictly be classified in such terms. As was noted earlier, a fundamental defining property of a multiple-baseline design is its staggered introduction of the intervention from one case's position to the next; and with Marascuilo and Busk's random of sampling of intervention start points for each case *with replacement*, it is possible that two or more cases will end up with the same intervention start point.

To rectify that problem, Levin et al. (2016) modified the procedure by restricting it to sampling *without replacement*, thereby assuring that no two cases will be assigned the same intervention start point. For Levin et al.'s restricted Marascuilo-Busk procedure, the researcher is also allowed to specify the minimum delay in the stagger desired between cases, ranging from an immediately adjacent stagger of one outcome observation to a function of the total number of outcome observations in the series. With N and k defined as above and s = the

minimum stagger desired between cases, the general formula for determining the total number of randomization-distribution outcomes is given by:

$$\frac{[k-(s-1)N+(s-1)]!}{[k-sN+(s-1)]!}$$

with the requirement that $k \geq s(N-1) + 1$. Note that with a desired minimum stagger of one observation (i.e., $s = 1$), for all values of N ($\leq k$), the formula reduces to:

$$k!/(k-N)! = k \times (k-1) \times (k-2) \times \ldots \times (k-N+1)$$

For the present example, then, with $N = 4$ cases and k = an 8-observation potential intervention start-point range for all cases, along with a specified minimum stagger of $s = 2$ observations for all cases, the total number of randomization-distribution outcomes would be reduced from 4,096 for the original Marascuilo-Busk procedure to [8 – (1 x 4) + 1]! / [8 – (2 x 4) + 1]! = 5!/1! = 120 for the restricted Marascuilo-Busk procedure according to the just-presented general formula. Routines for instantaneously providing randomly selected intervention start points with a specified stagger and for conducting the restricted Marascuilo-Busk randomization test may be found in Gafurov and Levin's (2016) *ExPRT* multiple-baseline program.

Koehler and Levin (1998) procedure. The Koehler-Levin procedure combines Wampold and Worsham's (1986) staggered case randomization and Marascuilo and Busk's (1988) intervention start-point randomization properties. In fact, the Wampold-Worsham represents a special-situation application of the Koehler-Levin procedure, in which there is only a single fixed staggered intervention start point (i.e., $k = 1$) for all cases. Similar to the Marascuilo-Busk procedure, a potential staggered intervention start-point interval is specified for each case's position within the multiple-baseline design, resulting in the random selection of a start point from each of N nonoverlapping intervals of size k. At the same time, and as with the Marascuilo-Busk procedure, here it is possible and permissible to designate different-sized intervals, k_1, $k_2,\ldots,k_N$ for the N different cases (see Footnote 2).[3] With equal-sized intervals for all cases, the total number of randomization-distribution outcomes is given by $N! \times k^N$. Thus, for a design

based on 4 cases and an intervention start-point interval of three observations for each case, that total would be equal to $4! \times 3^4 = 24 \times 81 = 1{,}944$.

Between-Case Randomization-Test Procedures

Modified Revusky procedure with either: (a) a single fixed intervention start point or (b) multiple potential intervention start points (Levin et al., 2016). Revusky's (1967) procedure is the earliest statistically sound nonparametric approach for analyzing the data from single-case multiple-baseline designs. Although the procedure is similar to the preceding within-case procedures with respect to determining each case's mean difference between the B intervention phase and the A baseline phase, it differs in terms of how it then examines those differences. Specifically: (1) starting with the first multiple-baseline position (referred to as Step 1), sequentially within each stagger position the mean difference associated with case for whom the intervention was just introduced is compared with the mean differences of lower-position cases for whom the intervention has yet to be introduced; (2) the mean differences at each step are rank ordered from 1 (largest) to *N*, *N-1*, etc., with the rationale that the case for whom the intervention was just introduced will produce a B-A mean difference larger than those for whom the intervention has not yet been introduced; and (3) the across-steps sum of the ranks is taken and, with the case-randomization assumption that cases were randomly assigned to the *N* multiple-baseline positions, that sum placed within a randomization distribution that contains the *N*! possible sums. For our example, with *N* = 4 cases, the total number of randomization-distribution outcomes is equal to 4! = 24, just as it was for the Wampold-Worsham test.

In an effort to make the original Revusky (1967) procedure more powerful, Levin et al. (2016) modified it in two distinct ways: first, by amalgamating the *N* steps' mean differences and applying a single rank ordering to those differences; and second, by adding intervention start-point randomization to Revusky's model. With only a single fixed intervention point for each multiple-baseline position, the original and modified Revusky procedures differ only with respect to summing within-step B-A mean differences (original Revusky) as opposed to rank ordering

the complete set (i.e., combined across steps) of B-A mean differences (modified Revusky). With a staggered interval of *k* potential start points for each case, the total number of randomization-distribution outcomes is equal to $N! \times k^N$, just as it was for the earlier discussed Koehler-Levin procedure. Consequently, for our present example with $N = 4$ and $k = 3$, that number is 1,944.

A Note on Series Lengths

An additional characteristic that differentiates among the five multiple-baseline randomization-test procedures is pertinent and needs to be mentioned: namely, the series lengths that are required for the *N cases.* For the Wampold-Worsham and Koehler-Levin procedures, data for all cases must continue to be collected for at least as long as the final-position case's initial post-intervention outcome observation. So, for example, if the intervention for the last case is introduced just prior to Observation 15, then all of the preceding cases' B intervention phases must continue at least through Observation 15.

In contrast, with the restricted Marascuilo-Busk and modified Revusky procedures, the series-length requirements are different. For the restricted Marascuilo-Busk procedure, the only requirement is that each case is associated with at least one post-intervention outcome observation. For the modified Revusky procedure, on each step the lower-position cases' series must continue for at least as long as that of the higher-position case's potential intervention start-point interval. So, for example, if the potential start-point intervals were Observations 4-6, 7-9, 10-12, and 13-15 for Cases 1-4, respectively, then at Step 1, all four cases would need to continue at least through Observation 6; at Step 2, Cases 2-4 would all need to continue at least through Observation 9 (whereas Case 1's data collection could be terminated before that, if either desired or necessary); at Step 3, Cases 3 and 4 would need to continue at least through Observation 12; and at Step 4, Case 4 would need to continue at least through Observation 15.

Rationale for the Present Study

The present two-investigation simulation study represents a logical sequel to a recently reported three-investigation study (Levin et al., 2016). In that study it was found that: (1) all seven of the multiple-baseline randomization-test procedures that were compared—Wampold-Worsham, Koehler-Levin (two variations), restricted Marascuilo-Busk, and modified Revusky (three variations)—maintained their Type I error probabilities at acceptable levels; (2) the randomly selected intervention start-point models (Koehler-Levin, restricted Marascuilo-Busk, and modified Revusky) typically outperformed Wampold and Worsham's and the modified Revusky single fixed start-point procedures in terms of their statistical powers, with the restricted Marascuilo-Busk and Koehler-Levin procedures generally emerging as the most powerful; and (3) in situations where it is not possible (or desirable) to include comparable series lengths for all cases (which, as was just noted, is required for the Wampold-Worsham and Koehler-Levin procedures), the restricted Marascuilo-Busk and modified Revusky procedures were recommended as reasonably powerful methods of choice.

As was indicated by Levin, Ferron, and Gafurov (2014) and Levin et al. (2016), however, in almost all previous single-case randomization-test simulation investigations of which we are aware, the Type I errors and powers of various randomization tests have been compared in situations where the intervention effect was modeled to be immediate and abrupt. Illustrated in Panel A of Figure 1 is a prototypical immediate abrupt intervention effect is one in which the benefits of the intervention (represented by either an outcome increase, as in Figure 1, or a complementary decrease) are produced concurrently with the introduction of the intervention and remain at a constant level as long as the intervention continues to be administered. In the present study, and as was considered in an earlier study by Lall and Levin (2004), we extend these comparisons to situations in which the effect is modeled to be either delayed and abrupt (i.e., when the effect of the intervention is not produced until more than one observation following the intervention's administration) or immediate and gradual (i.e., when the benefits of the intervention increase with continued administration of the intervention)—see Figure 1's

Panel B for a one-observation-delayed abrupt effect and Panel C for an immediate gradual effect.

In the present Investigation 1, we investigate the extent to which Levin et al.'s (2016) seven different multiple-baseline procedures are sensitive to the two just-mentioned alternative effect types, delayed abrupt and immediate gradual. Then, upon observing the statistical power losses produced by the alternative effect types, in Investigation 2 we explore "ameliorative actions" (i.e., modifications and, hopefully, improvements of Investigation 1's randomization-test procedures) that are available to single-case researchers who are still seeking to implement sensible and sensitive multiple-baseline randomization design-and-analysis strategies.

Investigation 1

Method

We considered multiple-baseline designs with the number of cases being four or five, and the series lengths being 19 for the $N = 4$ cases and 22 for the $N = 5$ cases. For the two single fixed intervention start-point designs, we included the Wampold-Worsham (WW/3) and modified Revusky (Rev-M/3) procedures, where the "/3" indicates a between-case stagger of three observations. A stagger of three observations for these two procedures was dictated by Levin et al.'s (2016, Investigation 1) findings that providing a between-case stagger of three observations increases the procedures' power beyond that which is attained from between-case staggers of either one or two observations. For the five multiple potential intervention start-point designs, we included the Koehler-Levin [KL(2) and KL(3)], modified Revusky [Rev-M(2) and Rev-M(3)], and restricted Marascuilo-Busk (MB-R/1) procedures, where the "(2)" and "(3)" respectively indicate two and three potential intervention start points for each case and the "/1" indicates a between-case stagger of at least one observation.

The intervention start points were set in advance to 6, 9, 12, and 15 for the $N = 4$ designs and 6, 9, 12, 15, and 18 for the $N = 5$ designs. With the Koehler-Levin and modified Revusky procedures, for the $N = 4$ design and two potential intervention start points per case,

the start points were randomly selected from {6, 7}, {9, 10}, {12, 13} and {15, 16}. For $N = 5$, we added the potential start-point set of {18, 19}. With the restricted Marascuilo-Busk procedure, for the $N = 4$ design the interventions were chosen without replacement from the interval 6 to 15, inclusive, and for the $N = 5$ design from the interval 6 to 18, inclusive.

Time-series data were generated for each case using error generation methods that paralleled those of Levin et al. (2016). A normally distributed error series was created for each participant using SAS's autoregressive moving-average simulation function (ARMASIM) within SAS IML (SAS, 2013). We originally considered including error series autocorrelations of $\rho = 0$ and .30, based on our wanting both: (*a*) to examine values above and below the average bias-adjusted autocorrelation of .20 that was found in a survey of single-case studies (Shadish & Sullivan, 2011); and (*b*) to use values that had been used in previous simulations of multiple-baseline data (e.g., Ferron & Sentovich, 2002; Ferron & Ware, 1995; Levin et al., 2016). However, because Levin et al.'s Investigation 1 results documented that: (1) with an autocorrelation of $\rho = 0$, Type I error is well controlled for all sample sizes; (2) although the powers were uniformly higher for $\rho = 0$ than for $\rho = .30$, the respective test-procedure profiles are virtually the same for both levels of autocorrelation; and (3) additional simulations confirmed those findings; in the present study, we elected to examine only the more realistic single-case research autocorrelation situations, namely, those for which $\rho = .30$.

In all simulations the standard deviation of the independent (i.e., uncorrelated) portion of the error series was set to 1.0. The observed data for our simulated studies was created by adding the error series created using the ARMASIM function to an effect-size series, which was coded to have values of 0 for all baseline observations and values of *d* for all intervention phase observations. Specifically, and as was indicated by Levin et al. (2016, p. 9):

> …the intervention effect was simulated as [a] change in level by the amount *d*, where *d* is defined as the difference between the case's B and A phase mean parameters, divided by the standard deviation of the independent portion of the error series. The value of *d* was varied to examine both Type I error control (when $d = 0$) and power (*d*s ranging from .5 to 4 in increments of .5). The choice

> of *d* values was informed by a survey of single-case intervention studies reported by Parker and Vannest (2009), where the estimated values of *d* (assuming no autocorrelation for simplicity and expected maximum powers) for the 10th, 50th, and 90th percentiles were 0.46, 1.70, and 3.88, respectively.

As we have mentioned in the past (e.g., Levin et al., 2012), *d*-defined effects that have been reported in various single-case intervention research literatures are generally much larger in magnitude than what Cohen (1988) initially regarded as "large" (i.e., *d* = .80) in the traditional psychological research literature. In particular, *d* values of 1.0, 1.5, and 2.0 exceed Cohen's (1988) "large" effect-size adjectival rule of thumb for traditional between-subjects designs but were considered here because it has been previously noted that *d* effect sizes of 2.0 or more are not uncommon in empirical single-case intervention research. In fact, Rogers and Graham (2008, p. 885) uncovered numerous effect sizes in excess of 3.0 in their meta-analysis of single-case writing interventions.

In contrast to the Levin et al. (2016) study, here the change in level of interest is not immediate and abrupt, but a change that was modeled to represent one of the three alternative effect types. For immediate gradual effects, the effect series was coded to have values of 0 for the elements corresponding to baseline observations, 1/3 *d* for the element corresponding to the first intervention observation, 2/3 *d* for the element corresponding to the second intervention observation, and *d* for all remaining elements in the effect series. We also examined delayed abrupt effects that were delayed by either one post-intervention observation or two post-intervention observations. For the former, the effect series was coded as 0 for elements corresponding to baseline observations, 0 for the first element during intervention, and *d* for all remaining elements in the effect series. For the latter, the effect series was coded as 0 for elements corresponding to baseline observations, 0 for the first two elements during intervention, and *d* for all remaining elements in the effect series. As was noted above, the observed sequence was obtained by adding the error series to the *d* effect series.

By crossing the number of cases (N = 4, 5) by the effect size (d = 0 to 4, in increments of .5), by the type of effect (immediate gradual, delayed-by-one-observation abrupt, or delayed-by-two-observations abrupt), 54 unique conditions were formed. For each of these conditions, data for 10,000 “studies” were simulated for each type of randomized design and then analyzed with the corresponding randomization test. More specifically, for designs with a random ordering of cases, a single fixed start point, and a between-case stagger of three observations, the WW/3 and Rev-M/3 tests were conducted; for designs with a random ordering of cases and a random selection of two [or three] intervention start points the KL(2) and Rev-M(2) [or KL(3) and Rev-M(3)] tests were conducted, and for designs where the intervention start points were randomly selected (without replacement) from specified intervals with a between-case stagger of at least one observation, the MB-R/1 test was conducted. For each test the obtained test statistic was based on the mean difference between intervention and baseline observations. The Type I error (when d = 0) or power (when d > 0) was estimated for each test for each of the 162 conditions by determining the proportion of the 10,000 simulated studies where the test led to a one-tailed p-value of .05 or less.

Results and Discussion

The results are presented in Figures 2 and 3, with the two leftmost panels (A1 and B1) of Figure 2 based on N = 4 cases and those of Figure 3 based on N = 5 cases. The different test procedures being compared are, in the A panels: Wampold-Worsham with a between-case stagger of three observations (WW/3), Koehler-Levin with either two [KL(2)] or three [KL(3)] potential intervention start points per case, and the restricted Marascuilo-Busk procedure with a between-case stagger of at least one observation (MB-R/1); and in the B panels, the modified Revusky procedure with one fixed intervention start point per case and a between-case stagger of three observations [Rev-M/3] (which is equivalent in power to the original Revusky procedure in the N = 4 design—Levin et al., 2016, Investigation 1) and the modified Revusky procedure with either two or three potential intervention start points per case [Rev-M(2) and Rev-M(3),

respectively]. The rejection rate of the null hypothesis is shown as a function of the effect size for designs having an autocorrelation of .30. Figure 2 is based on *N* = 4 cases with 19 observations per case, where the first actual or potential intervention start points for the four cases are 6, 9, 12, and 15 for all procedures except MB-R/1, which had start points selected at random without replacement from the interval 6 to 15 inclusive. Figure 3 is based on *N* = 5 cases with 22 observations per case, where the first actual or potential intervention start points for the five cases are 6, 9, 12, 15, and 18 for all procedures except MB-R/1, which had start points selected at random without replacement from the interval 6 to 18 inclusive.

Presented in both figures are effects that were: immediate and abrupt (A1/B1), from Levin et al. (2016); immediate and gradual (A2/B2); delayed by one observation and abrupt; (A3/B3), and delayed by two observations and abrupt (A4/B4). To reduce graph clutter, the two previously discussed test-procedure classifications, within-case and between-case, have been separated in Figures 2 and 3, with the former presented in the A panels and the latter in the B panels. In addition, for ease of comparison, the actual numerical highest and lowest power values for effect sizes .5 to 2.0 (along with other selected effect sizes) have been printed in each graph.

*Immediate abrupt effects (*from *Levin et al., 2016)*. In Panels A1 and B1 of Figures 2 and 3, it may be seen that for both sample sizes all seven test procedures exhibited satisfactory Type I error control (i.e., average values less than or equal to .05 when *d* = 0), with mean empirical αs ranging from .042 to .058. The only instance of an average empirical α in excess of .053 occurred with the Rev-M(3) procedure based on *N* = 4 cases. Concerning power (when *d* > 0), three aspects of Figures 2 and 3 are worth noting. First, of the within-case procedures depicted in Panel A1, there was little difference in power among the three procedures with a randomly selected intervention start point [KL(2), KL(3), and MB-R/1], each of which was superior to the single fixed intervention start-point Wampold-Worsham procedure with a

between-case stagger of three observations (WW/3). The power differences are substantial for $N = 4$ cases and only modest for $N = 5$ cases.

Second, and paralleling the within-case procedure findings, the same was true for the two Panel B1 between-case procedures with a randomly selected intervention start point [Rev-M(2) and Rev-M(3)], relative to the fixed single start-point modified Revusky procedure with a between-case stagger of three observations (Rev-M/3)—again, substantially for $N = 4$ and modestly for $N = 5$. Third, from Figures 2 and 3 it is clear that the within-case procedures of Panel A1 are more powerful than their between-case counterparts in Panel B1. Specifically, the three former randomly selected intervention start-point procedures [KL(2), KL(3), and MB-R/1)] are superior to both of the latter procedures [Rev-M(2) and Rev-M(3)], as is the former single fixed start-point WW/3 model relative to the latter Rev-M/3 model. For $N = 4$, for example, there is a power difference of about 10 percentage points to detect an effect size of $d = 1.5$ favoring the within-case procedures over the corresponding between-case procedures.

Alternative effect types. Also presented in Figures 2 and 3 for $N = 4$ and $N = 5$ cases, respectively, are the Type I error and power for effects that are immediate and gradual (Panels A2 and B2), effects that are delayed by one observation and abrupt (Panels A3 and B3), and effects that are delayed by two observations and abrupt (Panels A4 and B4). It is important to note here that the Figure 2 and 3 *d* values greater than 0 on the X axis for the three alternative effect types (i.e., the *d*s entering into power calculations) cannot be interpreted in the same manner as the *d* values associated with the immediate abrupt intervention effects. That is because for immediate abrupt effects (Panels A1 and B1), *all* intervention (B-phase) observations were defined to yield an effect size of *d*, resulting in a mean intervention effect size of *d*. In contrast:

(a) For immediate gradual effects (Panels A2 and B2), the first two intervention observations were defined as 1/3 *d* and 2/3 *d*, respectively, with all subsequent intervention observations (*P*) equal to *d*. As a result, the operative effect size (*d**) ranges from 2*d*/3 for one

subsequent intervention observation (i.e., $P = 1$) to $(1 + P)d / (2 + P)$ for P subsequent intervention observations.

(b) For one-observation-delayed abrupt effects (Panels A3 and B3), the first intervention observation was defined as 0, with all subsequent intervention observations defined as d. Therefore, d^* ranges from $d/2$ for one subsequent intervention observation ($P = 1$) to $Pd / (1 + P)$ for P subsequent intervention observations.

(*c*) For two-observation-delayed abrupt effects (Panels A4 and B4), the first two intervention observations were defined as 0, with all subsequent intervention observations defined as d. Consequently, d^* ranges from $d/3$ for one subsequent intervention observation ($P = 1$) to $Pd / (2 + P)$ for P subsequent intervention observations.

For all three alternative effect types, then, the mean intervention effect size must be regarded as something less than the immediate abrupt effect-type's mean of d, with the amount less determined by the type of alternative effect examined and the number of intervention observations (P) associated with an effect size equal to d.

1. Type I error control. As in Levin et al.'s (2016) study, the Type I error control associated with the various randomization-test procedures was for the most part acceptable, with averages ranging from .039 to .058 for the $N = 4$ designs and from .038 to .056 for the $N = 5$ designs. Of the 42 situations (7 procedures by 3 effect types by 2 sample sizes) investigated, only four yielded average empirical Type I errors that were in excess of .055. All of these were associated with the modified Revusky procedure based on either two or three potential intervention start points.[4]

2. *Power*. First, in striking contrast to Levin et al.'s (2016) investigations' powers associated with immediate abrupt effects (re-presented here in Panels A1 and B1 of Figures 2 and 3), the powers associated with the present investigation's effect types take a serious "hit" for all procedures—the most dramatic of which are the powers for two-observations-delayed abrupt effects (Panels A4 and B4 of Figures 2 and 3).

Consider as an example the single fixed intervention start-point Wampold-Worsham procedure based on $N = 4$ cases with a between-case stagger of three observations (WW/3). Levin et al. (2016) found that the power was .66 to detect an immediate abrupt effect of $d = 1.5$ (see also Panel A1 of Figure 2). However, from Panels A2, A3, and A4 of Figure 2 it may be determined that the powers for that procedure are only .38, .38, and .13 to detect, respectively, a $d = 1.5$ immediate gradual effect, a one-observation-delayed abrupt effect, and a two-observations-delayed abrupt effect.[5] A comparable pattern is associated with the randomized intervention start-point extension of the Wampold-Worsham procedure, namely, Koehler and Levin's regulated randomization procedure. Levin et al. reported power of .79 for that procedure based on $N = 4$ cases and two potential intervention start points per case [KL(2)] to detect an immediate abrupt effect of 1.5 (see also Panel A1 of Figure 2), whereas from Panels A2, A3, and A4 of Figure 2 it may be determined that the powers are only .41, .34, and .12, to detect, respectively, the above alternative $d = 1.5$ effect types. A similar KL(2) precipitous drop in power occurs when $N = 5$, where they are .92, .65, .60, and .24, respectively. A comparison of Levin et al.'s results (re-presented in Panel A1 and B1 of Figures 2 and 3) with Panels A2/B2, A3/B3, and A4/B4 of the same figures (along with the later-presented Table 1) convincingly demonstrates that although most of the randomization-test powers investigated here have respectable power for detecting immediate abrupt effects of a moderate size, they are inadequate for detecting other effect types of the same magnitude. This conclusion should be taken especially seriously for effects that emerge in more than a one-observation-delayed (rather than in an immediate) fashion.

Furthermore, along with the power declines seen in Panels A2/B2, A3/B3, and A4/B4 of Figures 2 and 3 for all test procedures, differences among them are also apparent. Generally speaking, when an immediate gradual effect was specified (Panels A2 and B2 in Figures 2 and 3), the within-case procedures (Wampold-Worsham, Koehler-Levin, and restricted Marascuilo-Busk) fared better than the between-case modified Revusky procedures with either a single

fixed intervention start point or a randomly selected start point). Similarly, for abrupt effects delayed by one observation, and especially when *N* = 5 (Panels A3 and B3 in Figure 3), the within-case procedures (headed by Wampold-Worsham) were more powerful than the between-case procedures, with the modified Revusky procedure based on two potential intervention start points being distinctly inferior. Finally, for abrupt effects delayed by two observations, there was little to choose among the various procedures, as they were all lacking in power (Panels A4 and B4 in Figures 2 and 3). If a "best among the worst" choice had to be made for detecting such effects, however, it would be for the within-case Wampold-Worsham procedure when *N* = 5, even though for very large effect sizes (*d*s of 3.5 and 4) that procedure is actually surpassed by the between-case modified Revusky procedure with a single fixed intervention start point (Panel B4 in Figure 3).

Investigation 2

As was just noted, Levin et al. (2016) found that several randomization test procedures developed for the multiple-baseline design have reasonable power to detect intervention effects that are immediate and abrupt. However, the results of present Investigation 1 clearly indicate that for intervention effects that are either delayed or gradual, the power of these tests is seriously compromised. More generally stated, if a researcher is "off" in specifying the type of effect expected of the intervention (i.e., if there is a mismatch between the expected and obtained effect type), then none of the present randomization-test procedures—as formulated here—will do an adequate job. The "as formulated here" addition is critical however, in that it might be possible for a researcher to adapt some of the test procedures to account for different effect types. Potential "remedial" adaptations consist of adjusting the randomization-test procedure in anticipation of effects that are either delayed or gradual.

Delayed Abrupt Effects

One method of adjusting the Investigation 1 randomization tests (which were designed to detect immediate abrupt effects) is to redefine their test statistics so that they are better

aligned with the type of effect that is anticipated. Specifically, for effects that are anticipated to be delayed by one observation and abrupt, the test statistic could be defined as the difference between the A-phase mean and a B-phase mean that is calculated using all intervention observations except the first one. Similarly, for an anticipated effect that is delayed by two observations, the test statistic could be defined as the difference between the A-phase mean and a B-phase mean that are calculated using all intervention observations except the first two. We will refer to this test-statistic adjustment approach as the *data-omitting* method.

Another way that we considered for adjusting the randomization test in anticipation of effects that are delayed is to redefine the test statistic so that B-phase observations prior to the expected shift are considered as if they were A-phase observations. For effects that are anticipated to be delayed by one observation and abrupt the test statistic could be defined as the difference in the B- and A-phase means, where the B-phase mean is calculated using all B-phase observations except the first intervention observation and the A-phase mean is calculated using all the A-phase observations plus the first B-phase observation. Similarly, for an anticipated effect that is delayed by two observations, the test statistic could be defined as the difference in the B- and A-phase means, where the B-phase mean is calculated using all intervention observations except the first two and the A-phase mean is calculated using all the A-phase observations and the first two B-phase observations. With this method (which we will refer to as the *data-shifting* method), the randomization distribution is then formed by calculating the test statistic for all possible assignments designated within the randomization scheme.

With the Koehler-Levin procedure based on two potential intervention start points for all cases as an example, we now illustrate how the data-shifting procedure would be applied to detect an anticipated delayed-by-one abrupt effect. Suppose that the two potential start points for Case 1 are 6 and 7. Before the study begins, the researcher randomly determines whether the intervention will start in Session 6 or in Session 7, with an outcome observation taken in every session. If Session 6 is selected as the initial intervention (B-phase) observation, based

on the anticipated one-observation-delayed effect, it is nonetheless treated as the last baseline (A-phase) observation, while Session 7 is treated as both the first potential and the first actual intervention observation. Accordingly, if the predicted delayed-by-one abrupt effect materializes, the randomization test statistic should be the largest with that particular data division. Similarly, if Session 7 had been selected, Session 7 would be treated as the last baseline observation, with Session 8 treated as both the second potential and the first actual intervention observation associated with that start point. If the predicted effect were to materialize, it should produce the largest randomization test statistic with that particular data division. The process continues in this fashion for each of the subsequent staggered cases.

To apply the data-shifting adjusted test procedure to an anticipated two-observations-delayed abrupt effect, a data-analyst would proceed similarly, as follows: “Pretend” that the first potential intervention start point and observation occur not concurrently with the intervention’s actual introduction (here, Session 6) but two observations following it (Session 8), thereby shifting both the first potential intervention start point and the first actual intervention observation to that position and shortening the actual intervention series by two observations. Along with that, the data analyst would “pretend” that the baseline series extends two observations beyond the actual final baseline observation (from Session 1 through Session 7), thereby increasing the actual baseline series by two observations. The process would continue in the same fashion for the second potential intervention start point and the initial actual observation associated with that start point (i.e., starting in Session 9).

Although it is clear that the randomization distributions for the data-omitting and the data-shifting adjusted test procedures will differ, it is not clear to what extent these differences will lead to differences in statistical power and/or Type I error control. However, because the data-shifting method makes use of all the data for each randomization outcome whereas the data-omitting method does not, it might be suspected that the former procedure will exhibit

power advantages over the latter when the anticipated effect is properly specified and incorporated into the test statistic.

Immediate Gradual Effects

Suppose now that the effect is anticipated to be immediate and gradual, such that the anticipated effect of the first observation after intervention is 1/3 of the final effect, the anticipated effect for the second observation after intervention is 2/3 of the final effect, and the anticipated effect for all other B phase observations is 100% of the final effect. Accordingly, the test statistic could be calculated as the difference between the weighted B-phase mean and the A-phase mean, where weights of 1/3, 2/3, and 1 are given respectively to the first, second, and remaining B-phase observations. With this method, the test statistic that is chosen is calculated for the actual assignment and then compared to the distribution formed by calculating the test statistic for all possible assignments designated within the randomization scheme. We will refer to this as the *data-weighting* adjusted-test method.

In Investigation 2 we examine the comparative Type I error and power characteristics of these data-adjusting methods as applied to alternative effect types.

Method

The number of cases in the multiple-baseline design was set to four, with the designs and potential intervention start points being identical to the $N = 4$ multiple-baseline design from Investigation 1. There were several reasons for restricting the number of cases investigated here to four. First, rather than the investigation being comprehensive and exhaustive, we wanted it simply to serve as an illustration of how an adjusted randomization-test procedure can be formulated and implemented. Second, $N = 4$ multiple-baseline designs are commonly applied by single-case researchers. Finally, Levin et al.'s (2016) study revealed that when the intervention effects were immediate and abrupt, the differences between the approaches based on random intervention start points and those that were not were more substantial with $N = 4$ than with $N = 5$ cases.

For each case, the approach to generating data was identical to the approach used in Investigation 1, where an error series (which was generated with an autocorrelation set so that ρ = 0 or .30) was added to an effect series (which was generated based on the same values for immediate gradual, delayed-by-one observation abrupt, and delayed-by-two observations abrupt effects as in Investigation 1). As was noted above, to serve as an illustration of the various adjusted randomization-test procedures (as opposed to a comprehensive power investigation), we focused selectively on just two values of *d*, 0 (to examine Type I error) and 1.5 (to examine power). Crossing the seven randomization test types [WW/3, KL(2), KL(3), MB-R/1, Rev-M/3, Rev-M(2), Rev-M(3)] with the three effect types, two effect sizes, and two levels of autocorrelation yielded 84 data conditions. For each of these conditions, 10,000 data sets were simulated. Each simulated data set with delayed effects was analyzed in two different ways (namely, by means of the data-omitting and the data-shifting approaches), whereas each data set with immediate gradual effects was analyzed by means of the data-weighting approach. For purposes of comparison, the to-be-presented tables of results also include the Type I error and power values that were obtained in Investigation 1, for which the researcher did not anticipate an effect other than an immediate abrupt one.

Results and Discussion

The simulated Type I error (when *d* = 0) and power (when *d* = 1.5) results associated with the present example's *N* = 4 multiple-baseline design with 19 outcome observations, based on a one-tailed Type I error probability of .05, are presented in Tables 1 and 2 for autocorrelations of 0 and .30, respectively. The first thing to note when comparing these two tables is the uniformly higher power values when the observations are not autocorrelated (Table 1) than when the autocorrelation is .30 (Table 2), thereby substantiating prior claims that autocorrelation is a critical issue to consider with respect to the statistical analyses of single-case data (for recent discussion of that issue, see Kratochwill & Levin, 2014; and Shadish, 2014).

Delayed abrupt effects. For each randomization-test procedure, the first row of Tables 1 and 2 (Original) displays the empirical Type I errors and powers associated with detecting immediate abrupt (IA), delayed-by-one abrupt (D1A), and delayed-by-two abrupt (D2A) effects with the original procedures that were constructed to be sensitive to immediate abrupt effects. So, for instance, in Table 2 consider the Koehler-Levin procedure with two potential intervention start points [KL(2)] and an autocorrelation of .30. When the immediate-abrupt effect model was applied to an immediate abrupt outcome, the average empirical Type I error was an acceptable .048 and the power to detect an effect of d = 1.5 was .792. When the same model was applied to delayed-by-one abrupt and delayed-by-two abrupt outcomes, although the empirical average Type I errors still maintained their nominal levels (.049 and .049, respectively), the powers were severely deflated (.339 and .118, respectively).

Now consider the data-omitting and data-shifting adjusted-test methods, which were designed to anticipate the alternative effect types. Specifically, when the data-omitting method was applied to the preceding KL(2) example in Table 2, the average empirical Type I errors were again quite acceptable (.048, and .046 for delayed-by-one abrupt and delayed-by-two abrupt outcomes, respectively), but more importantly, the empirical powers experienced the hoped-for increases (.660, and .584, respectively). Even more dramatic, however, were the data-shifting adjusted-test method's power results. With KL(2) empirical Type I errors of .049 and .050 for delayed-by-one and delayed-by-two abrupt effects, respectively, the corresponding powers were .783 and .751, which are not much lower than the .792 value when the original KL(2) procedure was applied to immediate abrupt effects. In addition, note that the same pattern of power differences among the randomization-test procedures is maintained when the adjusted-test methods are applied to delayed abrupt effects as was previously described when the original procedures were applied to immediate abrupt effects, namely: (1) with the data-shifting method, the two within-case procedures with randomly selected intervention start points (Koehler-Levin and restricted Marascuilo-Busk) outperform the modified Revusky between-case

procedure with either a single fixed or a randomly selected intervention start point; and (2) with both the data-omitting and data-shifting methods, the Koehler-Levin and restricted Marascuilo-Busk procedures exhibit noticeably higher power than the single fixed intervention start-point Wampold-Worsham procedure.

As an important aside, the data-omitting adjusted method cannot be incorporated into the modified Revusky procedure because of the procedure's unique stepwise approach to including or excluding case observations. Specifically, because the modified Revusky procedure uses only a portion of the B phase to define the B minus A phase differences (i.e., the portion of the target case's B phase that overlaps with the A phases of the yet-to-be-treated cases), the data-omitting adjusted method will leave no B observations to compute the B minus A phase difference for some possible assignments.

So, to summarize the results so far, for the two delayed abrupt effect types it is clear that both adjusted-test procedures (and particularly the data-shifting method) are able to recoup a good portion of the power losses that were experienced when the original randomization-test procedures were applied to effect types that are abrupt following a delay of one or two observations.

Immediate gradual effects. The situation is not so sanguine, however, when the present data-weighting adjusted-test method is applied to immediate gradual effects, presented in the two IG columns of Table 3 (left for $\rho = 0$, right for $\rho = .30$). It is true that the method's empirical Type I errors are satisfactory and its associated powers exhibit increases relative to those obtained when the original randomization-test procedures were used to detect immediate gradual effects. However, even this adjusted-test method comes nowhere near the power levels of the two just-discussed adjusted-test methods' ability to detect abrupt effects. At the same time, it is not difficult to speculate why, in a single-case randomization-test context, assessing gradual effects in the adjusted-test manner comes up short relative to the strategies we developed for assessing abrupt effects. Specifically, in order for the data-weighting method to

achieve its maximum potential power, each of the weights assigned to each intervention observation must correspond closely to the weights associated with the pattern of each case's actual first few individual intervention outcome observations. Given the random error component of the data-generation process (and in real-world applications, the unreliability of the outcome measure), this is highly unlikely to occur. In contrast, the random-error situation is not as critical for the data-omitting and data-shifting methods when applied to delayed abrupt effects. That is because at the point in time that the effect-size values ultimately change from 0 to *d*, they do so in one large increment, thereby resulting in observations close to those respective values that would be expected to overcome the random error component of the process.

Reconsideration of adjusted-test procedures for delayed abrupt effects. It should be apparent that a single-case researcher "loses" (with respect to power) if he or she applies an adjusted-test procedure to an anticipated delayed abrupt effect (*viz*., the data-omitting or data-shifting method) but (a) the effect turns out to an immediate abrupt one; or (b) if an abrupt effect occurs following a number of delayed observations that differs from what was anticipated by the researcher (e.g., an abrupt effect emerges after a two-observations delay rather than after a predicted a one-observation delay). On the other hand, applying either of those adjusted methods provides a nice balance to what would happen to power if an anticipated immediate abrupt effect becomes an abrupt effect that is delayed by one or two observations. In that regard, a reasonable question for a single-case researcher to ask is "Which is the better adjusted-test method to select for detecting abrupt effects, data-omitting or data-shifting?" There is no universally correct answer to this question, as a number of specific multiple-baseline design considerations would likely need to be taken into account (e.g., the number of cases, the total number of observations, the number of baseline and intervention observations, the number of potential intervention start points). Nonetheless, we can provide a few additional considerations that might be useful for single-case researchers to ponder.

First, we remind the reader that our foregoing observations and conclusions are derived from a four-case, 19-observations multiple-baseline design, which may have implications for the two adjusted randomization-test methods being compared herein. Recall that the data-omitting method eliminates one or two observations when calculating the B-phase mean, whereas the data-shifting method includes those observations as part of the A-phase mean. The B-phase means include exactly the same observations with both methods and so they are the same. With the just-described process, and as was alluded to earlier, there are always fewer observations going into the A-phase means for the data-omitting adjusted-test method. This issue is especially relevant for the first case's A-phase mean, in that with the data-omitting method there would be 5 or 6 observations with one or two potential intervention start points, as applied to detecting either a one- or two-observations delayed abrupt intervention effect, respectively, whereas with the data-shifting method there would be 6-9 observations for the same situation. With all else being equal, then: (1) fewer observations will produce less stable/reliable means, and so on those grounds alone the data-shifting method might be favored; while also recognizing that (2) the data-shifting method's stability advantage will not be as critical for designs with a larger number of outcome observations.

Second, a general answer to the question can be stated in terms of comparative risks and rewards. Specifically, although we found in this study that the data-shifting adjusted-test strategy exhibited greater power than the data-omitting method when a delayed abrupt effect was both anticipated and produced, in situations where an anticipated delayed effect is instead manifested as an immediate abrupt effect, the data-omitting method offers greater power protection. Accordingly, the data-omitting method would be a safer choice in the face of uncertainty about the immediacy of the effect. That is because if a delayed abrupt effect is anticipated and instead an immediate abrupt effect is produced, the actual one or two initial intervention observations are omitted and will not serve to reduce the magnitude of the abrupt effect. In contrast, with the data-shifting method those one or two initial intervention

observations will be considered to be the final part of the baseline series, which, if an immediate abrupt effect materializes, will serve to inflate the baseline series artificially. On the other hand, the data-shifting method would be the more rewarding adjustment method of choice (in terms of power) when the researcher has much certainty that the intervention effect will be a delayed abrupt one.

These just-described cautionary comments based on risks vs. rewards are similar to those that can be made for nondirectional (two-tailed) vs. directional (one-tailed) statistical tests, omnibus tests vs. planned contrasts, including sharply focused "predicted pattern contrasts" (Levin, Marascuilo, & Hubert, 1978, pp. 183-187; Levin & Neumann, 1999), as well as various single-case hierarchical linear modeling (HLM) approaches (see, for example, Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014).

General Discussion

The various single-case randomization-test procedures investigated here (both the older and more recently developed ones) have almost invariably been applied to intervention effects that were modeled to be immediate and abrupt (for an exception, see Lall & Levin, 2004). With specific reference to multiple-baseline designs, Levin et al. (2016) recently reported that some of the more recently developed procedures, and notably those with randomly determined intervention start points (such as the Koehler-Levin and restricted Marascuilo-Busk procedures), have both acceptable Type I error control and adequate power (i.e., average power values of at least .70) for detecting "prototypical" (e,g., Horner & Odom, 2014) immediate abrupt intervention effects of the magnitude often observed in single-case intervention studies. In the present investigation, however, the powers for the seven test-procedure variations examined were drastically reduced for effect types that were not of the immediate abrupt variety, even though the within-case Koehler-Levin and restricted Marascuilo-Busk procedures—and especially the single fixed start-point Wampold-Worsham procedure—proved to be more powerful in detecting both abrupt one-observation-delayed and immediate gradual effects than

the between-case modified Revusky variations.[6] Fortunately, we also found that when effects other than immediate abrupt ones are anticipated on an *a priori* basis by single-case researchers, remedial actions can be taken in the form of applying specially adapted versions of the original randomization-test procedures.

Delayed Abrupt Effects

Specifically, such remedial actions were effective for detecting delayed abrupt effects with the data-omitting method and, especially, with the data-shifting method (see Tables 1 and 2). It is important to mention that those two methods for anticipating delayed abrupt effects are not the only approaches that could have been formulated. For example, and as part of the present investigation, an initially examined alternative way of modifying the test procedure is to adjust the data set by permanently deleting all intervention observations that occur in time prior to the anticipated effect (which can be regarded as a data-*deleting* method). With that approach, if the researcher anticipated that the effect would be delayed by one observation and abrupt, the first intervention (B-phase) observation for each case would be removed from the data set. The altered data set would then be analyzed using a randomization test with a traditional test statistic based on the difference between the A- and reduced B-phase means. For effects that are anticipated to be abrupt and delayed by two observations, a similar approach would be followed, whereby the first two intervention observations of each case are deleted from the data set and then a randomization test is conducted using a traditional test statistic based on the difference between the A- and reduced B-phase means. As had been hoped for, this data-deleting method was found to produce a substantial power increase relative to the original unadjusted randomization test. However, the gain in power was accompanied by somewhat inflated empirical Type I errors, namely, around .06. A potential solution to the problem would be for the researcher to compensate for the data-deleting method's anticipated Type I error inflation by performing the test with a more conservative nominal Type I error probability ($\alpha = .04$, for example). Some preliminary simulations that we have conducted based on $\alpha = .04$ suggest that

this might represent a viable strategy, maintaining the empirical Type I errors in the .05 range but (as would be expected) reducing powers somewhat relative to the same tests conducted using $\alpha = .05$.

As a relevant aside, similar Type I error inflation (i.e., the tests were too "liberal") was reported by Lall and Levin (2004) in their simulation-study investigation of Levin and Wampold's (1999) independent start-point randomization-test model for comparing the effectiveness of two different between-case interventions. At the same time, Levin et al. (2012) found that certain ABAB…AB randomization-test procedures produced empirical Type I errors that were far lower than their nominal values (i.e., the tests were too "conservative") when the ABAB...AB sequence was based on "systematic" alternation as opposed to "randomized" alternation within pairs of observations. Unfortunately, no across-the-board advice can be given to single-case researchers about what to expect in the way of Type I error liberalism or conservatism concerning randomization tests whenever the tests that are performed do not meet Edgington's (1980) requirement for the tests to be "valid" (i.e., the type of randomization employed in the analysis is completely consistent with the type of randomization that was incorporated into the design). The newly proposed data-adjusted tests proposed here do not satisfy the Edgington requirement, which is why it was a necessity for each of their associated Type I error rates to be examined empirically.

Immediate Gradual Effects

Of the three alternative effect types examined here, the most challenging was an immediate gradual effect. Although the data-weighting test-adjustment method provided some compensation for the losses, the resulting power was still unacceptably low even for an eventual effect size of 1.5, as may be seen in Table 3. We note that in the present investigation all immediate gradual effects were of a specific form (i.e., 1/3 *d*, 2/3 *d*, *d*, *d*, *d*, …, *d*). Other types of gradual linear effects (such as 1/10 *d*, 2/10 *d*, 3/10 *d*, 4/10 *d*, etc.) and nonlinear gradual effects (e.g., 0.1*d*, 0.3*d*, 0.6*d*, *d*) have yet to be examined.

Practical Matters and Concluding Comments

Consideration is now given to the manner in which our data-adjusting techniques can be incorporated into the randomization-test analyses of real-world datasets. In particular, Gafurov and Levin's (2016) *ExPRT* package is capable of analyzing for anticipated delayed abrupt effects based on the present procedures. In particular, for the data-shifting method, program users simply specify their predicted number of delayed observations and the adjusted-test procedure is implemented directly. With that method, the B-A phase-mean difference for each case is shifted forward to begin with the B-phase's second or third outcome observation for one-observation and two-observations delayed effects, respectively, with all preceding observations regarded as part of the A phase (whether or not they were actual A-phase observations). In addition, the designated potential intervention start-point interval is shifted forward by the same amount (for this example, by two or three observations, respectively). Alternatively, to target a delayed abrupt effect with the data-omitting adjusted-test method, program scripts, such as those provided in R (Bulté & Onghena, 2009) or SAS (Levin et al., 2016), can be adjusted so that one or two B phase observations are omitted from the computation of the B phase means.

Similarly, for situations in which an immediate gradual A-to-B-phase change in level is expected (as was modeled here), the test-statistic computation can be adjusted in existing scripts such that the B-phase mean is a weighted mean. The Appendix provides an example of such an adjustment to an SAS script for the restricted Marascuilo-Busk procedure for which the test statistic has been adjusted by applying weights of 1/3, 2/3, and 1 to successive outcome measures of the B phase. In that regard, we re-emphasize that to benefit from any of the present randomization-test procedures, the researcher must have reasonably good knowledge about the type of intervention effect to be expected. Otherwise, there will likely be a mismatch between the type of effect programmed for and the actual effect produced, which, as was noted earlier, will result in a loss of power for detecting the latter.

The latter point suggests the need to provide additional guidance concerning which types of effects are likely to be anticipated for different outcome measures or in different contexts, from which they would be able to plan their analyses accordingly. Immediate gradual effects might be expected for interventions targeting the acquisition of an academic skill such as reading fluency, whereas delayed abrupt effects might be expected for many behavioral interventions. Although that might often be the case, the present authors do not believe that such characterizations are so cut and dried. One could imagine, for example, an intervention strategy for mastering a complex mathematics operation that takes time to be internalized and expressed but then manifests itself precipitously (a delayed abrupt effect); or the introduction of a potent positive reinforcer or form of punishment that, as soon as it is introduced, has a dramatic impact on a child's aberrant behavior (an immediate abrupt effect). The bottom line is that we are reluctant to provide blanket recommendations in the form of: "If you are intending to change skill/behavior X with intervention Y, then you should expect to produce an effect of type Z." Rather, interventionists must have a solid understanding—through personal experience with, literature reviews of, and/or extensive pilot research on—the specific substantive problem, subject population, targeted skill or behavior, intervention being applied, outcome measure(s), etc., in order to make educated guesses about the type of intervention effects that are likely to emerge.

This brings us to a related, crucial ethical caveat, about which the present authors have strong feelings. There can be no "do overs" when it comes to randomizing single-case intervention design components (interventions, intervention orders, cases, intervention start points, among others), lest the probabilisitic basis of randomization be undermined, thereby leaving questionable the researcher's claims about the intervention's impact (Ferron & Levin, 2014, Footnote 2; Levin et al., 2016, p. 18). Similarly, the study affords researchers the opportunity to select from seven different multiple-baseline randomization-test variations, while at the same time specifying the type of effect that they expect to emerge (namely, immediate

abrupt, immediate gradual, and delayed abrupt). Our clear intent is that the researcher should select both the specific randomization test to be applied and the type of effect to be expected strictly on an *a priori* (i.e., pre-experimental) basis. We are relatively hopeful that this process will be followed judiciously with respect to the particular randomization test selected because most of the tests are associated with uniquely different decisions about the nature and number of intervention start points that must be determined in the design phase of the study.

Specifying different effect types is another matter, however, as that process serves to test a researcher's integrity. Thus, if the researcher does not specify a particular anticipated effect type on an *a priori* basis (and, particularly, prior to examining the data), but rather conducts multiple analyses on the same data with different effect-type specifications, then we would again have ethical concerns and would question the validity of the researcher's statistical conclusions. This multiple hypothesis testing approach (variously known as "data dredging," "probability compounding," or the "multiplicity" problem— see, for example, Levin, 2006; and Maxwell, 2004) might be played out in the present multiple-baseline context as follows: The researcher conducts a randomization test, assuming a standard immediate abrupt effect, and which does not produce a statistically significant intervention effect. However, following the analysis the researcher notices that in each tier of the multiple-baseline there appears to be a two-observations-delayed effect and so, accordingly, the researcher reruns the same analysis with a two-observations-delayed effect specified and the test result is statistically significant. Following that type of procedure is akin to *post hoc* "model fitting," for which the researcher's stated Type I error probability associated with the analysis (of, say $\alpha = .05$) is no longer correct but instead is meaningless or misleading at best and fraudulent at worst.

We hasten to remind the reader that assessing changes in level (mean) from Phase A to Phase B is not the "only game in town" for single-case intervention researchers. A focus on between-phase changes in trend (slope) or variability is also a relevant concern (Horner & Odom, 2014). Research now in progress is examining whether the present multiple-baseline

randomization tests (including their adjusted-test versions) can be shaped to assess changes in trend and variability and, if so, to examine their comparative power characteristics.

With respect to the criticality of intervention start-point randomization for the procedures examined here, some traditional single-case researchers will argue that in multiple-baseline designs in particular, with the exception of the first participant, it is not appropriate to start the intervention at randomly selected points. That is because changes in the outcome variable for one participant cannot be convincingly demonstrated unless intervention implementation is withheld for subsequent participants. Moreover, the argument continues that introducing the intervention to a subsequent participant prior to demonstrating intervention effects for a participant who has already started the intervention would likely prevent publication in most journals that accept single-case research.

We are not claiming that the present randomized start-point procedures should be endorsed by or are suitable for all single-case interventionists, or that they should be applied in all single-case intervention research contexts. Response-guided single-case designs—where intervention start-point decisions depend on the emerging data—have a long history of use in school psychology, provide strong arguments for internal validity (e.g., Kazdin, 2011; Sidman, 1960), and can be conducted in a way that facilitates analyses that control Type I errors (Ferron & Jones, 2006) and that are sensitive to treatment effects (Joo & Ferron, 2015). However, school psychology researchers sometimes find themselves in contexts where, because of practical constraints, the study design has to be planned in advance, such as when the school calendar limits the number of weeks available for the study or when consent of school or clinic administrators depends on identification of the weeks that the intervention will be implemented. In addition, in situations where both response-guided and randomized designs are feasible, one researcher may choose a response-guided design based on arguments supporting those designs (for example, see Ferron & Jones, 2006; Kazdin, 2011; and Sidman, 1960), whereas another researcher may choose a randomized design based on arguments supporting those

designs (for example, see Edgington, 1996; Ferron & Levin, 2014). Because randomized designs are increasingly being considered and adopted by single-case interventionists (e.g., Ainsworth et al., 2014; Altschaefl, 2015; Bardon et al., 2008; Bice-Urbach & Kratochwill, 2016; Hwang et al., 2016; Markham et al., 2011; Onghena, Vlaeyen, & de Jong, 2007), we believe it important to study the Type I error control and power of analyses that have been proposed for those designs and design variations.

Finally, it can be argued that the present results and conclusions are based on a computer-simulation study, that single-case intervention researchers are seeking methodological and statistical strategies that are applicable to “real live” people, and that the twain might never meet. Of course that concern can be raised for any simulation study and, as a result, guardians of the academic research literature are always in the position of having to decide whether simulated outcomes are potentially relevant to the needs of their journal readership or not. We cannot defend the direct correspondence of the present findings and those that are characteristic of sentient human beings, other than to argue that: (1) the process we adopted to generate data used parameter values that were motivated by the results of analyses of time-series data from hundreds of single-case study participants (e.g., Parker & Vannest, 2009; Shadish & Sullian, 2011); (2) the impetus for our study came from the recognition that the effects that are produced by an intervention can be more complex (e.g., delayed or gradual) than what has been studied in previous simulation research; (3) the single-case intervention-produced level changes that are reported in the educational and psychological literatures typically resemble the three prototypical patterns that are portrayed in Figure 1; and (4) to the extent that they do, our results would have at least some modicum of real-world applicability. It should also be noted that for a variety of reasons, interventions that do not produce statistically reliable effects supporting the researcher’s preferred hypothesis (i.e., interventions that yield null effects/negative results) typically do not appear in either the traditional group research or single-case research literatures, resulting in a systematic

publication bias that unfortunately overestimates the true magnitude of the intervention effect (e.g., Ferguson & Heene, 2012; Greenwald, 1975; Shadish, Zelinsky, Vevea, & Kratochwill, 2016).

In sum, many novel variations and promising new directions are currently being carved out for the application of single-case randomization tests. These versatile methodological and data-analysis tools should therefore be welcomed by school psychology researchers insofar as they provide those researchers with a viable option for conducting scientifically credible single-case intervention investigations.

References

Ainsworth, M. K., Evmenova, A. S., Behrmann, M., & Jerome, M. (2016). Teaching phonics to groups of middle school students with autism, intellectual disabilities and complex disabilities needs. *Research in Developmental Disabilities*, *56*, 165-176.

Altschaefl, M. R. (2015). *Promoting treatment integrity of parent- and teacher-delivered math fluency interventions: An adult behavior change intervention*. Unpublished doctoral dissertation, University of Wisconsin, Madison.

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.

Barton, E. E., & Reichow, B. (2012). Guidelines for graphing data with Microsoft® Office 2007™, Office 2010™, and Office **f**or Mac™ 2008 and 2011. *Journal of Early Intervention*, *34*, 129-150.

Bice-Urbach, B. J., & Kratochwill, T. R. (2016). Teleconsultation: The use of technology to improve evidence-based practices in rural communities. *Journal of School Psychology*, *56*, 27-43.

Brewer, N., White, J. M. (1994). Computerized handwriting instruction with severely mentally handicapped adults. *Journal of Intellectual Disability Res*earch, *38*, 37-44.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods*, *41*, 477-485.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ: Erlbaum.

de Jong, J. R., Vangronsveld, K., Peters, M. L., Goossens, M. E. J. B., Onghena, P., Bulté, I., & Vlaeyen, J. W. S. (2008). Reduction of pain-elated fear and disability in post-traumatic neck pain: A replicated single-case experimental study of exposure in vivo. *Journal of Pain*, *9*, 1123-1134.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology*, *90*, 57-58.

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

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, *34*, 567-574.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561.

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: Statistical and methodological advances* (pp. 153-183). Washington, DC: American Psychological Association.

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

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

Gafurov, B. S., & Levin, J.R. (2016, Oct.). *ExPRT* (*Excel 7 Package of Randomization Tests*): *Statistical analyses of single-case intervention data* (Version 2.1). Downloadable from *http://ex-prt.weebly.com/*.

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

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27-51). Washington, DC: American Psychological Association.

Hwang, Y., Levin, J. R., & Johnson, E. W. (2016). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design**.** *Developmental Neurorehabilitation*. Online version available at http://dx.doi.org/10.3109/17518423.2015.1100689.

Joo, S.-h., & Ferron, J. M. (2015, November). *A simulation study of masked visual analysis of single case data*. Paper presented at the annual meeting of the Florida Educational Research Association, Altamonte Springs, FL.

Kazdin, A. E. (2011). *Single-Case Research Designs* (2nd ed.). New York: Oxford University Press.

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

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.

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. (1992). Single-case research design and analysis: Comments and concerns. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New developments for psychology and education* (pp. 213-224). Hillsdale, NJ: Erlbaum.

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

Levin, J. R. (2006). Probability and hypothesis testing. In J. L. Green, G. Camilli, & P. B. Elmore, (Eds.), *Handbook of complementary methods in education research* (pp. 519-537). Mahwah, NJ: Erlbaum.

Levin, J. R., Evmenova, A. S., & Gafurov, B. S. (2014). The single-case data-analysis *ExPRT* (*Excel*[7] *Package of Randomization Tests*). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp.185-219).

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; retrievable from http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2016). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*, Online version available at http://dx.doi.org/10.1080/17518423.2016.1197708.

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., 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., & Neumann, E. (1999). Testing for predicted patterns: When interest in the whole is greater than in some of its parts. *Psychological Methods*, *4*, 44-57.

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.

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

Markham, P. T., Porter, B. E., & Ball, J. D. (2011). Effectiveness of a program using a vehicle tracking system, incentives, and disincentives to reduce the speeding behavior of drivers with ADHD. *Journal of Attention Disorders*, 17, 233-248.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

McCleary, R., McDowall, D., & Bartos, B. J. (in press). *Design and analysis of time series experiments*. Oxford, UK: Oxford University Press.

Onghena, P., Vlaeyen, J. W. S.,& de Jong, J. (2007). Randomized replicated single-case experiments: Treatment of pain-related fear by graded exposure in vivo. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 387-396). Charlotte, NC: Information Age Publishing.

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

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*, *31*, 33-50.

Revusky, S. H. (1967). Some statistical treatments compatible with individual organism methodology. *Journal of the Experimental Analysis of Behavior*, *10*, 319-330.

Rogers, L. A., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, *100*, 879–906.

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

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122.

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

Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication preferences of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*. DOI: 10.1002/jaba.308.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Tyrrell, P. N., Corey, P. N., Feldman, B. M., & Silverman, E. D. (2013). Increased statistical power with combined independent randomization tests used with multiple-baseline design. *Journal of Clinical Epidemiology*, *66*, 291-294.

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

Footnotes

1. The question of whether the *N* cases' intervention start points must be staggered is considered in a subsequent section.
2. It is important to mention that with the Marascuilo-Busk procedure, different intervention start-point ranges can be specified for each case. If, in addition, all of the cases' specified start-point ranges are nonoverlapping, the procedure evolves into the next-discussed method proposed by Koehler and Levin (1998), for which case randomization is also included.
3. An alternative, more readily implemented "successively restricted" sampling procedure—not examined here and similar to a suggestion made by Koehler and Levin (1998, Footnote 7)—can be formulated to ensure more equal between-case staggers. With that procedure, the researcher would randomly select the first case's intervention start point from the usual pre-specified acceptable interval. After seeing where that start point falls, the researcher would specify a new acceptable interval from which the second case's start point is to be selected. That new range would be specified with the desired amount of stagger taken into consideration, and the same process would continue for the remaining cases.
4. Levin et al. (2016, Investigation 2) also found that even with an immediate abrupt change specified, for an autocorrelation of .30 and $N = 4$ cases the modified Revusky procedure based on three potential intervention start points produced an average empirical Type I error of .058.
5. It should be noted, however, that the single fixed start-point Wampold-Worsham procedure exhibits a nontrivial power advantage relative to its three randomized start-point competitors for both delayed abrupt effect types (Panels A3 and A4 of both figures), as does the single fixed start-point modified Revusky procedure (Rev-M/3)

relative to the two randomized start-point modified Revusky procedures Rev-M(2) and Rev-M(3) for all three alternative effect types (Panels B2, B3, and B4 of both figures).

6. Tyrrell, Corey, Feldman, and Silverman (2013) have recently proposed an alternative power-increasing strategy for the Wampold-Worsham procedure, based on combining the test results of two separately conducted $N = 4$ multiple-baseline studies. In addition, from a distinctly practical perspective, we remind the reader that of the various randomization-test approaches investigated here, with the restricted Marascuilo-Busk and modified Revusky procedures there are no constraints regarding the length of the cases' respective intervention series, whereas with the Wampold-Worsham and Koehler-Levin procedures there are (see also Levin et al., 2016).

Appendix

The following sample SAS program script conducts the randomization test for an immediate gradual effect for a four-case restricted Marascuilo-Busk design, for which the range of potential intervention start points for each case is from Observation 6 through Observation 15, with a minimum between-case stagger of one observation. The user would replace the x matrix values by entering his/her outcome data from each case as a row in x. The user would then alter the values of the initial potential start point (pi) and the final potential start point (pf) by entering his/her values in place of the current values of 6 and 15. The user would then enter the actual intervention start points for the four cases (ps1, ps2, ps3, and ps4, respectively) to reflect their values as opposed to the current values of 7, 9, 11, and 14, respectively. When the program is run, the *p*-value is printed. In this example the observed test statistic is the 23rd largest in the distribution of 5040 test statistic values, and thus the *p*-value is reported as .00456, which is 23/5040.

Program Script:

```
proc iml;
x={30 35 25 15 25 25 40 55 77 80 85 75 80 90 80 85 75 80 85,
   20 35 10 15 30 40 35 20 50 70 75 85 90 90 75 85 80 90 75,
   15 10  5  0 10 20 10 15 20 10 35 50 60 80 85 80 90 75 80,
   35 20 25 20 15 10 30 40 20 15 25 30 15 50 70 85 80 95 90};
pi=6; pf=15; ps1=7; ps2=9; ps3=11; ps4=14;
ncases=nrow(x);
nobs=ncol(x);
 a1o=sum(x[1,1:ps1-1])/(ps1-1);
 b1o=((1/3)*x[1,ps1]+(2/3)*x[1,ps1+1]+sum(x[1,ps1+2:nobs]))/(nobs-ps1);
 a2o=sum(x[2,1:ps2-1])/(ps2-1);
 b2o=((1/3)*x[2,ps2]+(2/3)*x[2,ps2+1]+sum(x[2,ps2+2:nobs]))/(nobs-ps2);
```

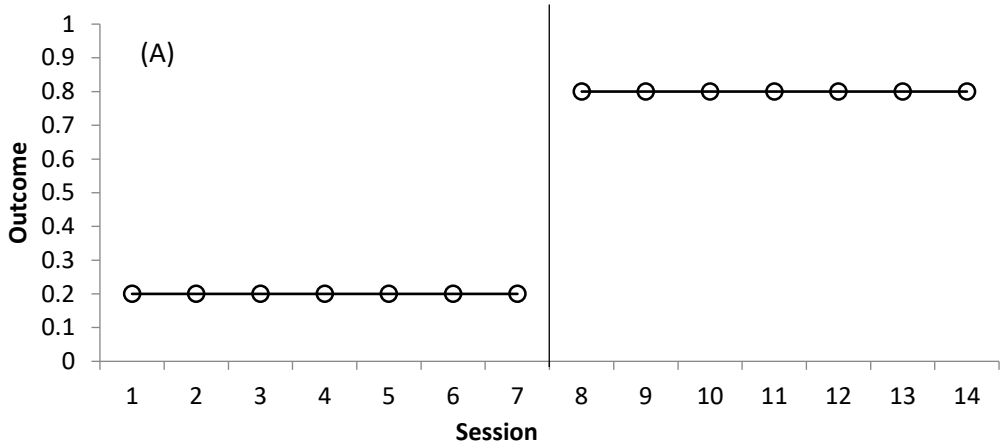
```
 a3o=sum(x[3,1:ps3-1])/(ps3-1);
 b3o=((1/3)*x[3,ps3]+(2/3)*x[3,ps3+1]+sum(x[3,ps3+2:nobs]))/(nobs-ps3);
 a4o=sum(x[4,1:ps4-1])/(ps4-1);
 b4o=((1/3)*x[4,ps4]+(2/3)*x[4,ps4+1]+sum(x[4,ps4+2:nobs]))/(nobs-ps4);
 test1ob=round(1000000*((b1o-a1o)+(b2o-a2o)+(b3o-a3o)+(b4o-a4o))/4);
counter1=0; count1=0;
do i=pi to pf;
do j=pi to pf; if j ^= i then do;
do k=pi to pf; if k ^= i & k ^= j then do;
do m=pi to pf; if m ^= i & m ^= j & m ^= k then do;
 counter1=counter1+1;
   a1=sum(x[1,1:i-1])/(i-1);
   b1=((1/3)*x[1,i]+(2/3)*x[1,i+1]+sum(x[1,i+2:nobs]))/(nobs-i);
   a2=sum(x[2,1:j-1])/(j-1);
   b2=((1/3)*x[2,j]+(2/3)*x[2,j+1]+sum(x[2,j+2:nobs]))/(nobs-j);
   a3=sum(x[3,1:k-1])/(k-1);
   b3=((1/3)*x[3,k]+(2/3)*x[3,k+1]+sum(x[3,k+2:nobs]))/(nobs-k);
   a4=sum(x[4,1:m-1])/(m-1);
   b4=((1/3)*x[4,m]+(2/3)*x[4,m+1]+sum(x[4,m+2:nobs]))/(nobs-m);
   testst1=round(1000000*((b1-a1)+(b2-a2)+(b3-a3)+(b4-a4))/4);
   if testst1>=test1ob then count1=count1+1;
   pvalue1=count1/counter1;
end; end; end; end; end; end; end;
print test1ob count1 counter1 pvalue1;
quit;
```

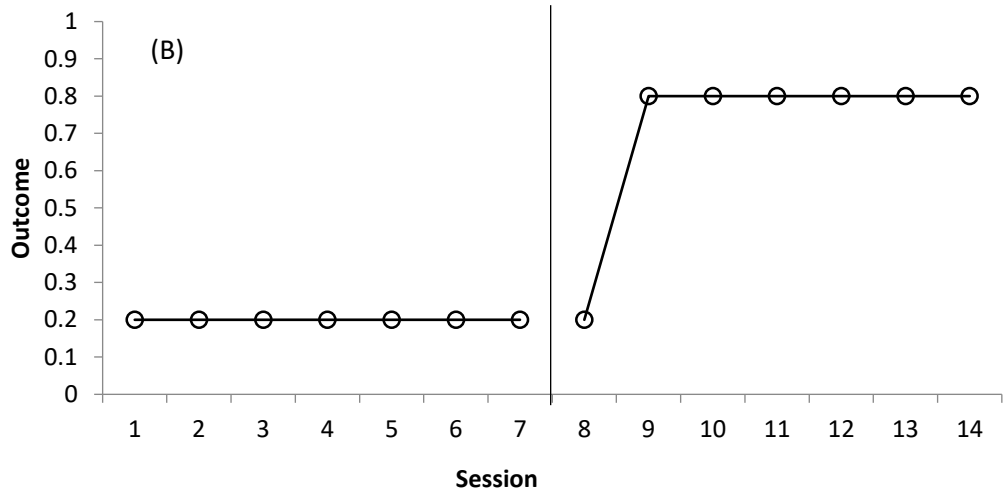
Figure Captions

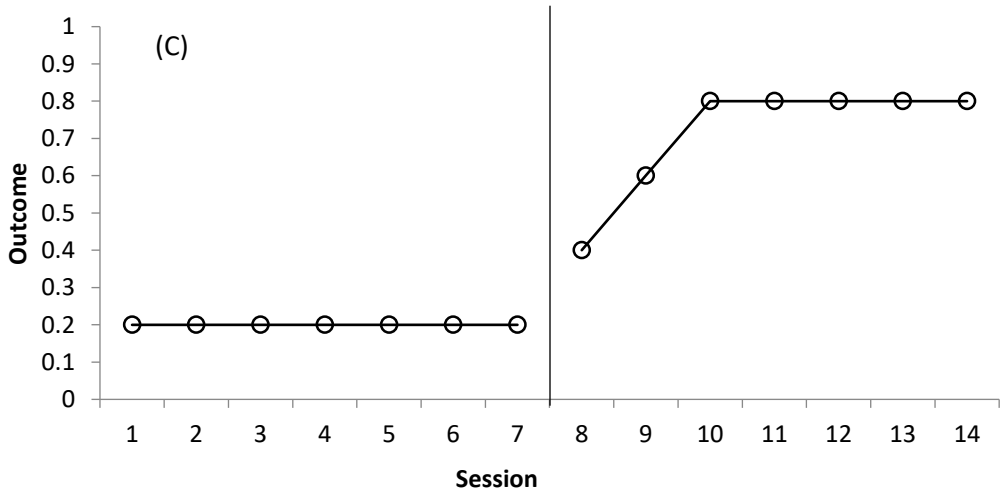
Figure 1. Prototypical representation of immediate abrupt (Panel A), one-observation-delayed abrupt (Panel B), and immediate gradual (Panel C) effect types

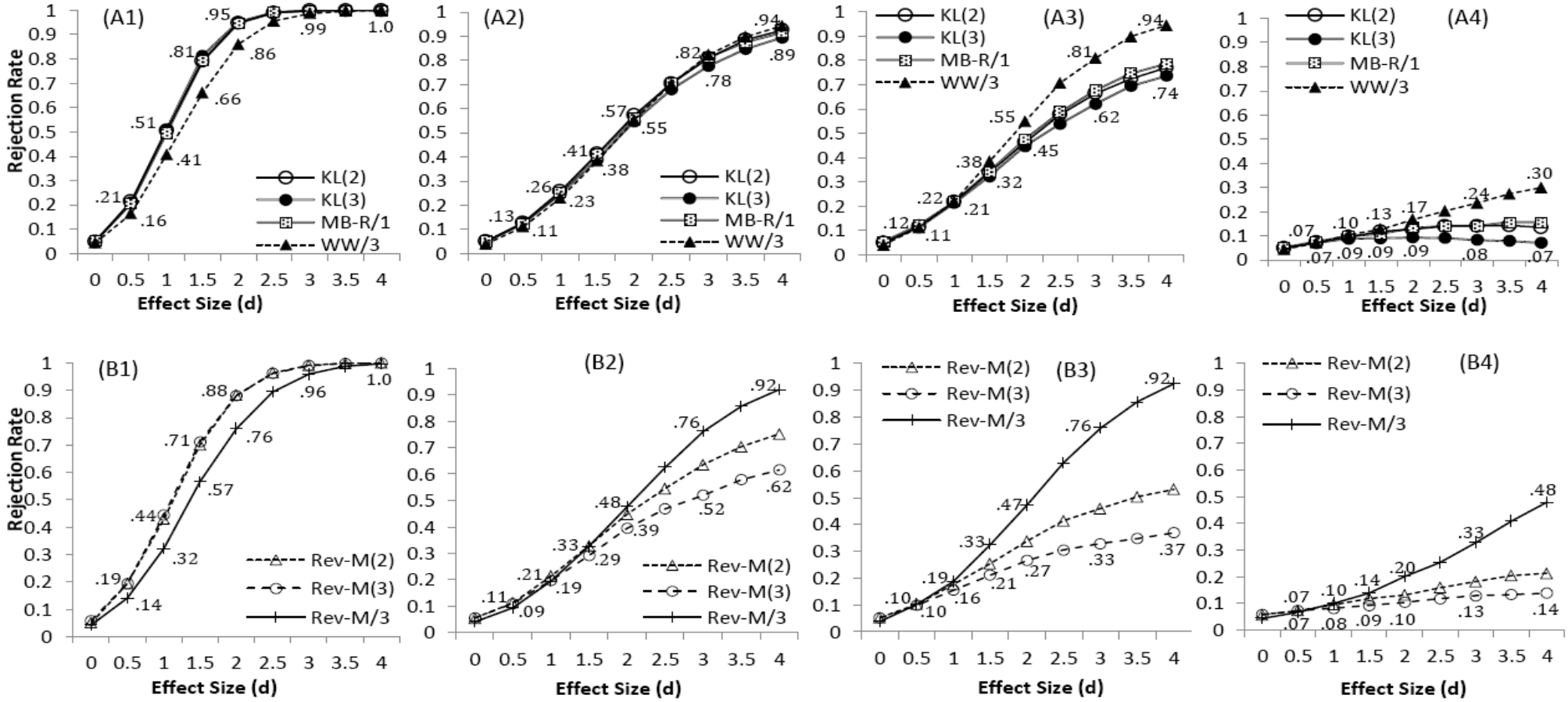
Figure 2. Power comparisons (α = .05, one-tailed) of the seven multiple-baseline randomization-test procedures/variations based on *N* = 4 cases to detect immediate abrupt effects (Panels A1 and B1), immediate gradual effects (Panels A2and B2), delayed-by-one observation abrupt effects (Panels A3 and B3), and delayed-by-two-observations abrupt effects (Panels A4 and B4)

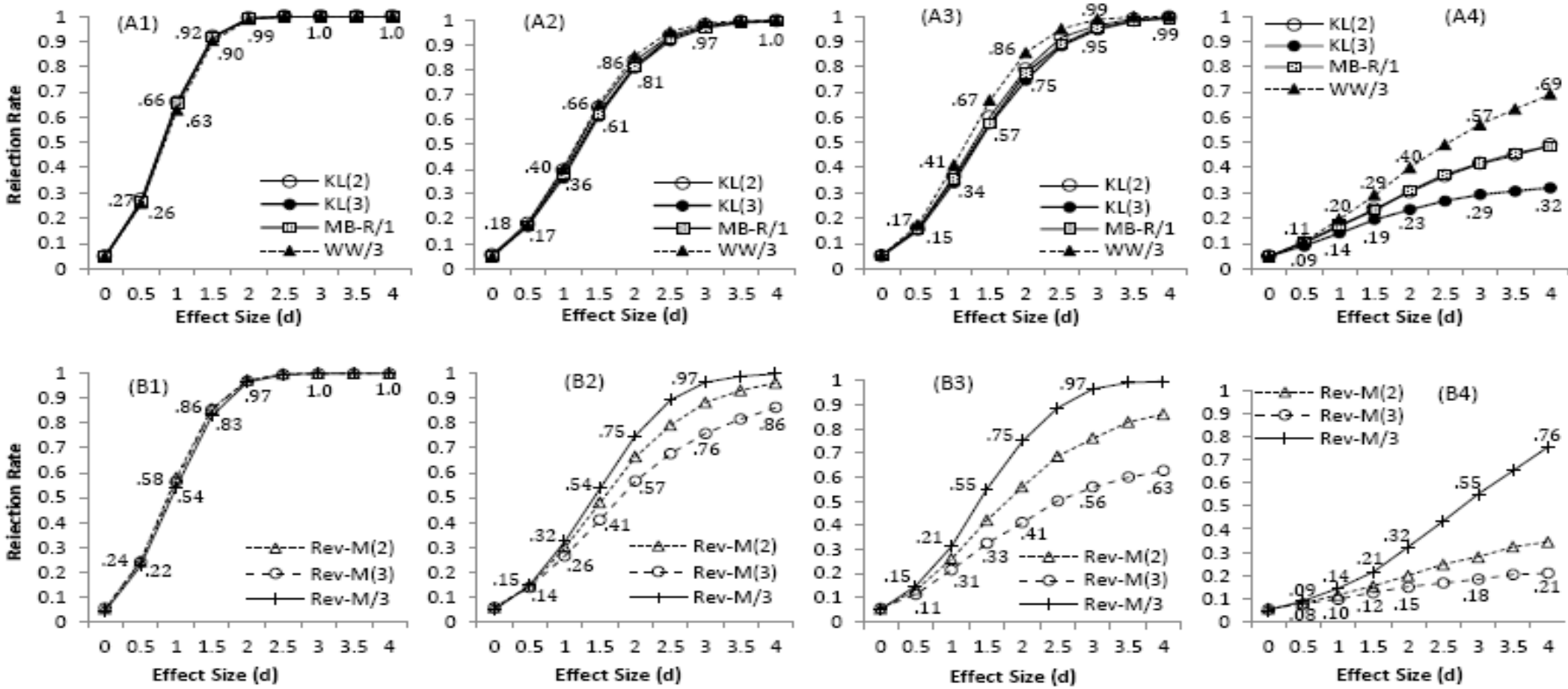
Figure 3. Power comparisons (α = .05, one-tailed) of the seven multiple-baseline randomization-test procedures/variations based on *N* = 5 cases to detect immediate abrupt effects (Panels A1 and B1), immediate gradual effects (Panels A2 and B2), delayed-by-one observation abrupt effects (Panels A3 and B3), and delayed-by-two-observations abrupt effects (Panels A4 and B4)


(A)
Outcome
Session
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14


(B)
Outcome
Session
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14


(C)
Outcome
Session
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14


(A1)
(A2)
(A3)
(A4)
(B1)
(B2)
(B3)
(B4)
Rejection Rate
Effect Size (d)
KL(2)
KL(3)
MB-R/1
WW/3
Rev-M(2)
Rev-M(3)
Rev-M/3


(A1)
(A2)
(A3)
(A4)
(B1)
(B2)
(B3)
(B4)
Reiection Rate
Effect Size (d)
KL(2)
KL(3)
MB-R/1
WW/3
Rev-M(2)
Rev-M(3)
Rev-M/3