Investigation of Single-Case Multiple-Baseline Randomization Tests of Trend and Variability

Joel R. Levin, University of Arizona

John M. Ferron, University of South Florida

Boris S. Gafurov, George Mason University

Abstract

Previous simulation studies of randomization tests applied in single-case educational intervention research contexts have typically focused on A-to-B phase changes in means/levels. In the present simulation study, we report the results of two multiple-baseline investigations, one targeting between-phase changes in slopes/trends and the other targeting between-phase changes in variability. For each of these measures, we examine the comparative Type I errors and powers of several randomization test procedures that have previously appeared in the literature. In so doing, we propose an alternative measure of variability that is more sensitive to detecting between-phase change than is the variance itself. We conclude by providing a summary table of recommended randomization test procedures for assessing different types of intervention-based effects associated with level, trend, and variability.

Investigation of Single-Case Multiple-Baseline

Randomization Tests of Trend and Variability

Single-case intervention educational research consists of carefully designed and implemented "time series experiments" (e.g., Glass, Gottman, & Wilson, 1976; McCleary, McDowall, & Bartos, in press) in which a series of temporally produced observations or measures is taken on one or a few "cases" (represented by individuals, group, or other aggregates) in a baseline or control phase (the A phase). An experimental treatment or intervention is then systematically introduced by an experimenter/researcher during which an additional series of observations is taken during the intervention or experimental phase (the B phase). To improve both the methodological rigor and outcome generalizability of the basic AB design, variations and extensions of it include: (1) replicating the A and B phases both within cases (e.g., an ABAB…AB design) and across cases, often in a staggered ("multiple-baseline") fashion (e.g., Horner & Odom , 2014; Maggin, Cook, & Cook, 2018; Plavnick & Ferreri, 2013); as well as (2) incorporating methodological procedures that are part and parcel of scientifically credible randomized "group" intervention research (Kratochwill & Levin, 2010; Levin, 1994). In addition, in recent years several innovative procedures have emerged for analyzing single-case educational intervention data, including those based on visual/graphical methods, statistical methods, and combinations of the two (see, for example, Busse, McGill, & Kennedy, 2015; Chen, Peng, & Chen, 2015; Ferron & Jones, 2006; Ferron, Joo, & Levin, 2017; Kratochwill & Levin, 2014; Manolov & Moeyaert, 2017).

In response to a reviewer's question about the tradeoffs between visual and statistical analyses of single-case data, we regard the two data-analysis approaches not as competitors but as complements. Visual analysis can be initially implemented to demonstrate "experimental control" and statistical analysis can be subsequently conducted on the data to provide quantitative confirmation. Moreover, as Hwang, Levin, and Johnson (2018, p. 233) state:

> The visual analyst arrives at a conclusion about the intervention's effectiveness primarily by taking into account the [individual participants'] individual responses to the intervention, whereas the statistical analyst relies primarily on an aggregated measure of the [participants'] outcomes…By considering the pluses and minuses associated with the individual cases, visual analysts formulate a conclusion about intervention effectiveness in the study as a whole… [R]andomization tests, on the other hand, are focused on aggregated summary measures, which will lead the analyst to the same statistical conclusion every time. Even in situations where not all (or even none) of the individual case profiles would document an intervention effect through visual analysis, a statistical analysis based on the combined cases might [and vice versa].

As far as an educational intervention researcher is concerned, the question arises: "Why adopt single-case rather than conventional "group" intervention research designs?" Here are a few reasons. First, single-case intervention investigations are valuable when resources are "scarce" as, for example, when the number of participants to be included in the study is limited because of economic factors or because the participants are to be selected from low-incidence populations (children with autism or students with severe reading disabilities) – keeping in mind, however, that relative to conventional group studies, far more outcome measures are typically required in single-case time-series studies. Second, in comparison to conventional group intervention research studies conducted in schools or in the community, single-case studies typically include fewer logistical hurdles with respect to obtaining permissions, coordination with school or community personnel, scheduling students, and other administrative requirements. Third, single-case research permits – indeed, strongly encourages as a *raison d'etre* – examination and in-depth analyses of individuals' performance. Fourth, single-case research serves well as a precursor or complement to a large-scale study.

Lessons learned, consistencies, and incongruities gained from single-case studies can serve larger-scale conventional group studies well. Finally, it should be recognized that single-case studies can afford intervention researchers attempts to replicate the results of conventional group studies, much in the way that effects obtained in traditional within-subjects designs allow for comparisons and contrasts with effects obtained in traditional between-subjects designs.

Newly developed single-case statistical-analysis procedures encompass quantitative non-overlap methods (e.g., Parker, Vannest, & Davis, 2014), single-level regression models (e.g., Beretvas & Chung, 2008; Maggin et al., 2011), multilevel modeling (e.g., Moeyaert et al., 2013; Rindskopf & Ferron, 2014), and Bayesian approaches (e.g., Shadish et al., 2014), among others. The present two-investigation Monte Carlo simulation study focuses on a different class of statistical-analysis strategy, namely, nonparametric randomization tests – an approach that is completely compatible with the various novel forms of randomization that have been proposed for single-case intervention research investigations (e.g., Heyvaert & Onghena, 2014; Kratochwill & Levin, 2010; see also Ferron & Levin, 2014, Levin, Kratochwill, & Ferron, 2019; and Michiels & Onghena, 2018). The applicability and versatility of randomization tests in single-case educational intervention research have been illustrated both in recent simulation studies (e.g., Bouwmeester & Jongerling, 2020; Levin, Ferron, & Gafurov, 2014, 2017ab, 2019; Levin, Ferron, & Kratochwill, 2012; Levin, Lall, & Kratochwill, 2011; Manolov, 2019; Michiels, Heyvaert, & Onghena, 2018), as well as in actual single-case research investigations (e.g., Ainsworth, Evmenova, Behrmann, & Jerome, 2016; Collier-Meek, Sanetti, Levin, Kratochwill, & Boyle, 2019; de Jong et al., 2008; Holden et al., 2002; Hwang & Levin, 2019; Hwang, Levin, & Johnson, 2018).

As Hwang and Levin (2019) note:

> What has elevated the scientific credibility of [single-case intervention] designs is their incorporation of various forms of randomization that characterize

> conventional experimental and randomized controlled trials research, incorporations serving to enhance the designs' internal validity…Within the single-case intervention design repertoire is the multiple-baseline design, which is among the most popular, versatile, and scientifically credible designs with respect to its ability to document a direct link between interventions and outcomes…Along with the recent methodological enhancements of single-case intervention designs, in general, and multiple-baseline designs, in particular, are concomitant enhancements in the statistical-conclusion validity of the data-analysis tools that are implemented to assess single-case research outcomes. (p. 161)

Randomization in Single-Case Educational Intervention Research

Different forms of randomization have been specified for various single-case educational intervention designs and their associated randomization statistical tests (e.g., Craig & Fisher, 2019; Edgington, 1996; Ferron & Levin, 2014; Jacobs, 2019; Levin, Ferron, & Gafurov, 2014; Levin, Kratochwill, & Ferron, 2019; Tanious & Onghena, 2019). These include the following, some of which are incorporated into the present simulation study:

1. within-case intervention-order randomization in AB designs, reversal designs (ABAB, in four adjacent phases: baseline – intervention – return to baseline – intervention) and alternating treatment designs(rapid alternation between experimental conditions: e.g., ABABABABAB); in certain applications of these designs, it is possible to randomize the order in which the multiple A and B phases are administered (e.g., BAABABABBA — see, for example, Levin et al., 2012).

2. between-case intervention (or intervention-order) randomization in replicated AB-type designs, such as when A and B represent two different interventions, with either the A or B intervention randomly assigned to cases (Levin, Ferron, & Gafurov, 2018) or both interventions

randomly assigned to cases in different orders (as in a crossover design format — see Levin, Ferron, & Gafurov, 2014).

3. random assignment of cases to the different stagger positions in multiple-baseline designs (e.g., Wampold & Worsham, 1986).

4. random assignment of intervention start points (or "transition points") to cases in all varieties of single-case intervention design, as was originally proposed by Edgington (1975) and then extended by Marascuilo and Busk (1988) — see Ferron & Levin (2014).

5. combinations of two or three of the preceding randomization forms (Levin et al., 2014).

Single-case educational intervention researchers should be familiar with the first three forms of randomization listed above: namely, randomizing the order in which interventions are administered in AB-type and alternating-treatment designs, both within and between cases (Nos. 1 and 2, respectively) and randomly assigning cases to stagger positions in multiple-baseline designs (No. 3). The value of such randomized single-case designs is that they improve the internal validity of an intervention study — for additional discussion, see Ferron & Levin (2014), Kratochwill & Levin (2010), and Levin, Kratochwill, & Ferron (2019). On the other hand, although the fourth form of randomization (intervention start-point randomization) was initially proposed five decades ago by Edgington (1975), and which also represents a critical internal-validity enhancer of single-case intervention research, it may not be as familiar a concept to single-case intervention researchers and so it will be briefly described here (see also Edgington, 1996; Heyvaert & Onghena, 2014; Levin et al., 2014; Manolov & Solanas, 2009; Michiels & Onghena, 2019).

Intervention Start-Point Randomization

With Edgington's (1975) intervention start-point randomization model, prior to data collection a researcher specifies, for a given case, a range of observations for which it would be "acceptable" to transition from the A baseline phase to the B intervention phase (i.e., "potential"

intervention start points). Once that designation has been made, the researcher randomly selects a single "actual" start point from within the range. Note that this *random* determination of the actual intervention start point is strikingly different from two other commonly applied intervention start-point determination strategies. Preferred by many traditional experimental methodologists is a *fixed* strategy, for which the researcher decides prior to data collection (based on previous research, preference, or necessity) the specific observation point at which the case is to transition from the baseline phase to the intervention phase. In contrast, preferred by traditional single-case intervention researchers is a *response-guided* strategy, for which the researcher closely follows the case's baseline series of responding as it occurs, with the intervention phase beginning only after a stable level of responding has been established. Additional discussion of these different intervention start-point options is provided by Ferron and Levin (2014, p. 160), Hwang et al. (2018), and Levin, Kratochwill, & Ferron (2019).

Adopting a random intervention start-point strategy (as is in the present simulation study) is advantageous from both a methodological and a statistical point of view. Concerning the former advantage, and as with the single fixed intervention start-point approach, a predetermined randomly selected intervention start point eliminates a major potential source of researcher bias (i.e., subjectivity) from the intervention-research process. That is, along with the other forms of single-case design randomization, incorporating intervention start-point randomization serves to render the study more scientifically credible.

As a relevant aside, the conceptual differences among these three different forms of intervention implementation (random, fixed, response-guided) have critical implications for single-case researchers that are similar to those associated with our earlier discussed visual and statistical analyses. Specifically, the notion of "experimental control" has distinctly different interpretations for the quantitative methodologist and the traditional behavior-analytic interventionist. For the methodologist, achieving experimental control through fixed or randomized intervention start-point administration connotes the same sense of scientific rigor as

is applied in conventional “group” randomized trials research, whereas for the interventionist, achieving experimental control through response-guided intervention administration implies capitalizing on the interventionist’s expertise concerning the specifics of when and how to introduce the intervention. A discussion of the pros and cons of each position is clearly beyond the scope of the present article and so for additional information the reader should refer to other sources (e.g., Gast, 2010; Kazdin, 2011; Kratochwill & Levin, 2010)

As for the statistical advantage of incorporating intervention start-point randomization, an associated randomization statistical test can provide a formal means of determining the degree to which any observed phase change (in, for example, mean/level) should be considered something other than a chance occurrence.[1] Specifically, the mean B-A (B minus A) phase outcome-data difference produced by the *actual* intervention start point for each case is calculated and combined with the mean B-A phase differences produced by the outcome data for all *potential* intervention start points, to form a randomization distribution. Then, the position of the actual mean difference within the randomization distribution is located, which, in turn can be converted to a significance probability. If the significance probability is less than or equal to the pre-established significance level (α), then the actual outcome is regarded as a statistically nonchance event. Complete details and examples of randomized intervention start-point statistical procedures for a variety of single-case intervention designs are provided by Marascuilo and Busk (1988), Koehler and Levin (1998), Levin, Ferron, & Gafurov (2014), Tanious & Onghena, 2019, among others.

Importantly, the statistical power of single-case randomization tests has been extensively examined in simulation studies and has been found to be respectable, especially with increases in the number of cases, as well as in the number of observations and potential intervention start points (e.g., Ferron & Sentovich, 2002; Levin et al., 2012; 2017a,b). When combined with other design-randomization strategies, such as randomizing the order in which

the A and B phases are administered in AB-type and alternating treatment designs, power increases can be substantial (Levin Ferron, & Gafurov, 2014).

Rationale for the Present Study

To date, simulation studies involving randomization statistical procedures have focused almost exclusively on tests of between-phase changes in level. Most recently, the present authors have examined the comparative statistical properties (notably, Type I errors and powers) of commonly adopted and newly developed randomization tests of level change in a variety of AB-type and multiple-baseline single-case intervention designs (Levin, Ferron, & Gafurov, 2014, 2017ab, 2019). As single-case methodologists and visual analysts hasten to point out, however, a change in level neither provides the complete picture of an "intervention effect" nor is it "the only game in town" (Horner & Odom, 2014; Tanious, De, & Onghena, 2019). Thus, here we focus on between-phase changes in "trend" (slope) and variability, as have been described by Levin, Evmenova, and Gafurov (2014). So, as one goes from Phase A to Phase B, we ask whether there is a change in the slope of the within-phase best-fitting straight line that predicts the case's outcome from the case's observation number (i.e., time period). That is, does the within-phase regression line become steeper or flatter, or does it change from positive to negative, between Phases A and B? Similarly, we ask whether the within-phase variability of the outcome measure increases or decreases between phases.

Consider, for example, the hypothetical A- and B-phase data in Figure 1. Panel 1 of Figure 1 was constructed to display a difference in slopes between the two phases. Specifically, the A-phase slope of .0167 is near zero, whereas the B-phase slope has increased to .2667. With these outcomes, the B-A phase difference in slopes is .25; or, in terms of the slope ratio, the B-phase slope is 16 times greater than the A-phase slope. This is apart from the B-A phase mean difference of 3.11 units. Similarly, the data in Panel 2 of Figure 1 were constructed to display differences in the two phase variances, which are 2.00 and .250 for Phases A and B,

respectively. These values produce an A-B phase variance difference of 1.75 and an A/B variance ratio of 8.00. The accompanying B-A phase mean difference is equal to .11 units. As far as changes in slope are concerned, in practice they are typically, though not always, accompanied by a a change in level, as when an intervention changes a stable (flat) baseline series to a gradual change in level during the intervention phase (as is illustrated in the next paragraph). However, a change in slope can also take the form of a stable baseline series transitioning during the intervention phase to an initial gradual increase in level followed by a return to a near-baseline level, as would occur with a temporary novelty or practice effect followed by an eventual habituation or fatigue effect. As for changes in variance (also illustrated in the next paragraph), these are often seen in practice either by: (a) floor-level baseline outcomes with little variability followed by considerably more variable responding once the administration of the intervention begins; or (b) considerable variability in baseline responding and little variability in responding following the administration of a highly effective intervention.

In the present two-investigation simulation study, we examine the Type I error and statistical power behaviors of various single-case randomization test procedures, which we apply here to the assessment of A-to-B phase changes in trend and variability. In comparing the different test procedures, we restrict our attention to multiple-baseline designs, all of which incorporate case randomization, with some also incorporating intervention start-point randomization. The present study was motivated by two primary considerations. First, randomization-based statistical tests of A-to-B phase changes in trend and variability have rarely (if ever) appeared in the single-case intervention literature.[2] Second, our focus on multiple-baseline designs and selected test procedures is a logical follow-up study to two recently conducted investigations in which the same design and test procedures were examined with respect to level changes, for both typically anticipated between-phase immediate abrupt changes and less typically anticipated between-phase delayed abrupt and immediate gradual changes (Levin et al., 2018). In the present context, immediate intervention effects refer to

between-phase changes that occur in the session immediately following the introduction of the intervention, whereas delayed intervention effects begin only two or more sessions following the intervention's introduction. Abrupt effects refer to between-phase changes that appear essentially "all at once" (i.e., in a single session) and remain at that level for several sessions, whereas gradual effects increase or decrease in magnitude as the sessions progress. These different prototypical effect types for changes in level are illustrated in Figure 2.

Here we investigate in depth the statistical sensitivity to between-phase trend and variability changes of five multiple-baseline randomization test approaches: (1) Wampold and Worsham's (1986) randomized case procedure (WW); (2) Koehler and Levin's (1988) dual case-randomization plus intervention-start-point randomization procedure, based on two potential intervention start points for each case [KL(2)], and which reduces to the WW procedure when a single fixed start point is established for each case; (3) Levin et al.'s (2018) restricted (*viz*., sampling without replacement) randomization multiple-baseline adaptation of Marascuilo and Busk's (1988) procedure (MB-R); (4) Levin et al.'s (2018) modified Revusky (1967) procedure based on one potential intervention start point for each case [Rev-M(1)]; and (5) Levin et al.'s (2018) modified Revusky procedure based on two potential intervention start points for each case [Rev-M(2)]. Because of the manner in which the test statistic is calculated (essentially, either case by case or across cases), the first three have been referred to as "within-case" procedures and the latter two as "between-case" procedures.

Each of these procedures follows the earlier discussed rationale for determining where the test statistic (based on a B-A phase difference in level, trend, or variability) that was *actually obtained* falls within a distribution consisting of all possible test statistics that *could have been obtained*, in accordance with the specific randomization-assignment process that was applied to cases, intervention start points, or both. The randomization distribution for each procedure is constructed in different ways. In brief: for WW, it is based on all possible permutations of the *N* cases. For KL(2), it is also based on permuting the *N* cases, and then combined with permuting

the *K* = 2 acceptable potential intervention start points for each case. For MB-R, it is based on the pre-experimental designation of a range of *K* acceptable potential nonoverlapping intervention start points for the *N* cases. Finally, for Rev-M(2), the randomization distribution is based on a set of sequential comparisons based on *N*, *N-1*, *N-2*,…,1 cases' B-A phase differences in conjunction with the *K* = 2 acceptable potential intervention start points for each case. Although the modified Revusky procedure has not fared well in previous simulated comparative power investigations, we include it here because there are certain design situations where it can be implemented when the Wampold-Worsham and Koehler-Levin procedures cannot be – namely, when it is not practically or economically feasible for all cases to complete all phases of the design (for specific restrictions, see Gafurov and Levin, 2019). The logic underlying each of these randomization-test procedures, as well as hypothetical examples providing their computational details, may be found in Levin et al. (2017b).

It is important to note that the Rev-M(2) simulations reported here are consistent with the way in which the procedure was explained and illustrated on pages 295-297 of Levin et al.'s (2017b) article. Undetected until recently is that a slightly different version of the Rev-M(2) procedure was implemented for the power simulations that were reported in the Levin et al. (2017ab) articles. Fortunately, however, based on subsequent simulations it was found that the obtained power differences between the two Rev-M(2) variations are negligible and so corrections of the previously reported findings are not warranted.

Investigation 1: Examination of Trend/Slope Randomization Tests

Method

We generated a time series for each case by adding a series of errors (e) to a series of true values (μ) such that at time *t* for case *i* the outcome value is $y_{ti} = \mu_{ti} + e_{ti.}$ We generated the errors for each time series using the autoregressive moving-average simulation function (ARMASIM) in SAS (2013) and specifying a first-order autoregressive model $e_t = \rho\, e_{t-1} + a_t$, where the variance of the white noise, VAR(a_t), was set to 1.0 and the autocorrelation, ρ, was

set to either 0.0 or 0.3. These values have been used in other simulations of multiple-baseline data (e.g., Ferron & Sentovich, 2002; Ferron & Ware, 1995; Levin et al., 2017b) and range from no autocorrelation (i.e., ρ = 0) to a value that is a little larger than the reported single-case meta-analytic average bias adjusted autocorrelation of .20 (Shadish & Sullivan, 2011). The true values were based on a stable baseline and an intervention effect that led to an immediate change in slope, which either extended throughout the intervention phase (no asymptote slope) or leveled off five observations into the intervention phase (asymptote slope).

The effect-size parameter, *d*, that was used to generate data was held constant across cases in a simulated study (i.e., $d = d_i$) and was operationalized as $d = \frac{\mu_{B5} - \mu_A}{\sigma_a}$, where μ_{B5} is the true value five observations into the intervention phase, $\mu_A$is the true value during baseline and σ_a is the standard deviation of the white noise in the first order autoregressive model. The relationship of the baseline standard deviation to the parameters of the first order autoregressive model is $\sigma_y^2 = \frac{\sigma_a^2}{1-\rho^2}$, where σ_a^2 is the variance of the white noise and ρ is the autocorrelation. Therefore, when the autocorrelation is 0 the effect size is indexed in baseline standard deviation units (i.e., the difference between the true value five observations into intervention and the true value during baseline equals *d* baseline standard deviations). For the conditions with an asymptote slope, *d* represented the final effect size.

The value of *d* was manipulated in the simulation to reflect different effect sizes ranging from *d* = 0 (i.e., no effect, to examine Type I error control) to *d* = 4 in increments of .5 (to examine power ranging from small to large effects). When the autocorrelation is 0 the effect size is indexed in baseline standard deviation units (i.e., the difference between the true value five observations into intervention and the true value during baseline equals *d* baseline standard deviations). The values of *d* were chosen to be in line with the effect estimates obtained in a survey of single-case intervention studies where the estimated values of *d* (assuming no autocorrelation and no trends) were 0.46, 1.70, and 3.88 for the 10th, 50th, and 90th percentiles of the distribution of observed effect sizes (Parker & Vannest, 2009). The series of true values

for the asymptote slope conditions was obtained by setting the baseline values to 0, the first four intervention observations to .2*d*, .4*d*, .6*d*, and .8*d* respectively, and all remaining intervention observations to 1.0*d* and, as a consequence, the difference between the baseline level and the asymptote in treatment ranged from 0 to 4 *SD*s.

The series of true values for the no asymptote slope conditions was obtained by setting the baseline values to 0, and each successive intervention-phase observation to .2*d* greater than the previous observation (i.e., .2*d*, .4*d*, .6*d*, .8*d*, 1.0*d*, 1.2*d*, etc.). Thus, when *d* was .5 the slope change was .1, which was the slope change studied by Joo, Ferron, Moeyaert, Beretvas, & Van den Noortgate (2019), and when *d* was 1 the slope change was .2, which was the slope change studied by Jamshidi et al. (2019). We also chose to include larger slope changes for two reasons: (1) we were simulating conditions where there was no immediate shift in level and thus larger slope changes were needed to get to the mean differences that had previously been studied; and (2) we anticipated that slope changes may be more difficult to detect than level changes and wanted our simulations to document how large the slope changes were required to be detected.

The time-series data were generated for multiple-baseline designs with five cases. The rationale for focusing on five-case designs, when four-case designs are more common in practice (e.g., Shadish & Sullivan, 2011), was that slope changes are more difficult to detect than level changes. We conducted similar simulations for the four-case multiple-baseline situation and all power estimates were unacceptably low, and thus rather than reporting those results here, we chose to focus on the results of the five-case multiple-baseline designs.

Each case was simulated to have 22 observations. The rationale for focusing on a series length of 22 is that it is close to the median series length of 20 reported by Shadish and Sullivan (2011) in their survey of single-case studies. By including two additional observations we were able to compare each of the randomization-test methods of present concern, including the Koehler-Levin and the modified Revusky procedure, each with two potential intervention start

points per case [KL(2) and Rev-M(2), respectively], while maintaining at least five baseline observations, and between-case staggers of at least 2 observations.

Intervention placement depended on the randomized design that was used. When either the Wampold-Worsham (WW) or the modified Revusky procedure with one potential intervention start point per case [Rev-M(1)] was used, the intervention start points for the five cases were 6, 9, 12, 15, and 18. When either the KL(2) or the Rev-M(2) procedure was used, the intervention start points were randomly selected from {6, 7}, {9, 10}, {12, 13}, {15, 16}, and {18, 19}. When the restricted Marascuilo-Busk (MB-R) procedure was used, the start points were chosen randomly without replacement from the interval 6 to 18 inclusive, based on a specified minimum between-case stagger of one observation.

By crossing the five randomization test methods [WW, KL(2), MB-R, Rev-M(1), and Rev-M(2)], by the effect size (d = 0 to 4, in increments of .5), by the two asymptote slope conditions (no asymptote slope versus asymptote slope), and by the level of autocorrelation (ρ = 0, .3), 180 conditions were formed. For each condition, data for 10,000 “studies”– resulting in a 95% confidence interval sampling error of less than .01– were simulated and analyzed. The data for each simulated study were then analyzed with a randomization test, where the permutations were based on the randomized design of the simulated study. The randomization test was carried out using as the test statistic (also referred to as the “measure”) the average difference between B- and A- phase slopes, $\Sigma(b_B - b_A)/N$.

Results and Discussion

The results are summarized in Table 1, where no asymptote is built into the B-phase slope. There it may be seen that when the A- and B-phase slopes are equal (i.e., when d = 0), the average rejection rates (i.e. the empirical Type I error probabilities) are well controlled for all the four procedures investigated, ranging from .032 to .047 when there is no autocorrelation in the series (i.e., when ρ = 0) and from .034 to .054 when the observations are moderately autocorrelated (i.e., ρ = .30). Consistent with previous multiple-baseline design randomization-

test investigations of level/mean differences (e.g., Levin et al., 2017ab), power decreases for all procedures [with the exception of Rev-M(2)] as the autocorrelation increases from 0 to .30 — here, sometimes by .10 or more power units. Concerning the comparative powers of the five procedures, the WW single fixed intervention start-point procedure and the procedure with the most potential intervention start points (MB-R) come out ahead, with the former slightly outperforming the latter; followed by KL(2); and with Rev-M(1) and Rev-M(2) lagging farther behind with clearly inadequate powers. For example, with a moderate autocorrelation of ρ = .30, powers for the largest effect size investigated (*d* = 4.0) are .75, .71, .65, .47, and .29 for WW, MB-R, KL(2), Rev-M(1), and Rev-M(2), respectively. Clearly, none of the procedures has adequate power to detect moderate slope-change effects: With ρ = .30 and *d* = 2.0, for example, the powers of all five procedures are less than .33.

The picture changes only slightly when an asymptote is built into the B-phase slope (see Table 2). Again, all five procedures adequately control their Type I error probabilities, with a range from .033 to .055 when the observations have no autocorrelation and from .035 to .052 when the autocorrelation is .30. The power profiles show WW, KL(2), and MB-R grouped somewhat closer together and all clearly superior to Rev-M(1) and Rev-M(2). With ρ = .30, for example, for the largest effect size of *d* = 4.0, the respective powers are .71, .67, .65, .46, and .28. Once again, however, when *d* = 2.0 all four procedures exhibit powers that are inadequate (*viz*., less than .33). A comparison of these asymptote slope powers with the previous no-asymptote slope powers reveals little difference between the two.

In sum, although the four single-case randomization test procedures originally developed to assess between-phase changes in level/mean are statistically valid to assess changes in trend/slope, even the three “leading” procedures for the latter proved not to be very powerful. This is illustrated from a simplified perspective in Figure 3 with respect to the WW perspective. In that figure is shown what three different *d* effect sizes correspond to when represented as *amounts* of slope change (i.e., when going from a baseline- phase slope of 0 to an intervention-

phase slope greater than 0). Specifically illustrated are effect sizes of d = 2.0, corresponding to an intervention-phase slope of .40 (Line a); d = 3.0, corresponding to an intervention-phase slope of .60 (Line b); and d = 4.0, corresponding to an intervention-phase slope of .80 (Line c). It is readily apparent that even with the largest slope change represented by Line c (a slope of 0 during the baseline phase and a no asymptote slope of .80 in the intervention phase), with 5 cases, 22 observations, an autocorrelation of .30, and a Type I error probability of .05, the WW procedure has only a 75% chance of detecting that change.

To determine whether the power situation would improve (and by how much) with increases in series lengths beyond 22 observations, we conducted a supplementary investigation.

Investigation 1a: Supplementary Examination of Trend/Slope Randomization Tests

To assess the impact of series length on power for detecting phase differences in slopes, we extended Investigation 1 based on five cases to include series lengths of 32 and 44 for two of the randomization test procedures [WW and KL(2)]. When series lengths were 32, the intervention start points for the WW method were 7, 12, 17, 22, and 27, and the intervention start points for the KL(2) method were randomly selected from {7, 8}, {12, 13}, {17, 18}, {22, 23}, and {27, 28}. When series lengths were 44, the intervention start points for the WW method were 12, 18, 24, 30, and 36, and the intervention start points for the KL(2) method were randomly selected from {12, 13}, {18, 19}, {24, 25}, {30, 31}, and {36, 37}. All other methods and conditions were the same as those of the primary study.

The results are summarized in Table 3 for both the no asymptote and asymptote slope situations. There it may be seen that in the no asymptote situation for both the WW and KL(2) procedures there are dramatic increases in power with increases in series length. With an autocorrelation of ρ = .30, for example, the WW powers associated with d = 2.0 are .33, .56, and .89, for series lengths of 22, 32, and 44 observations, respectively; and the corresponding KL(2) powers are .26, .53, and .86.

In striking contrast, in the asymptote slope situation, no such power increases with increasing series lengths are evident. For WW, the respective powers are .32, .30, and .28; and for KL(2) they are .29, .31, and .30. What is more, for the largest effect size of d = 4.0, there is an apparent *decrease* in power as the series lengths increase. The WW powers are .71, .63, and .53 for series lengths of 22, 32, and 44, respectively; and the corresponding KL(2) powers are .67, .64, and .54. The power decreases are understandable because following the initial nonzero slope values associated with the first five intervention-phase observations, all remaining intervention-phase observations were associated with a slope of zero. Consequently, with longer series lengths there will be more zero-slope observations included in calculating the complete intervention-phase slope, which serve to attenuate the slope based on the initial five intervention-phase observations.

Summary

Both good and bad news can be reported for the results reported so far. The bad news is that for all five of the multiple-baseline randomization tests of between-phase differences in slope/trend investigated, power is generally lacking when the design contains a typical number of observations (here, 22). The good news is that when the trend continues throughout the entire intervention-phase interval (i.e., when there is no slope asymptote) and the series lengths are increased to 32, and especially to 44, powers greatly increase. The same, however, cannot be concluded when the trend asymptotes after five intervention-phase observations. At the same time, we hasten to point out that although it is a simple matter to plug 32 or 44 outcome observations into a data-simulation exercise such as this one, including that many observations in an actual single-case intervention investigation may not represent a realistic situation. Shadish and Sullivan (2011) found the number of observations per participant in single-case studies was as high at 160, but the median value was 20 and 91 percent of single-case studies included fewer than 50 observations per case.

That said, we now turn our attention to randomization tests of changes from A- to B-phase variability in single-case multiple baseline designs.

Investigation 2: Examination of Variability Randomization Tests

Method

Consistent with the method used in the study of slopes, we generated a time series for each case by adding a series of errors (e) to a series of true values (μ) such that at time *t* for case *i* the outcome value is $y_{ti} = \mu_{ti} + e_{ti.}$ Again we generated the errors for each time series using the autoregressive moving-average simulation function (ARMASIM) in SAS and specifying a first-order autoregressive model $e_t = \rho\, e_{t-1} + a_t$ where the variance of the white noise, VAR(a_t), was set to 1.0 and the autocorrelation, ρ, was set to 0.0 or 0.3. Unlike the study for slopes, we modified the error values to produce differences in the error variance between phases. In particular, we multiplied the error of each intervention-phase observation by a constant, *R*, and consequently, the errors in the treatment phase had a *SD* that was *R* times the *SD* of the errors in the baseline phase.

Increasing *SD*s with treatment is common in behavior acquisition studies, whereas decreases in *SD*s are common in studies where the goal is to extinguish problem behaviors. We looked only at increasing SDs because we assumed the power would be comparable for decreasing *SD*s. The values of *R* were varied across conditions from a low of 1 (homogeneity) to a high of 8 (severe heterogeneity), in increments of 1. The range from 1 to 8 was motivated by a study of a reading instruction intervention for students with emotional and behavioral disorders (Barton-Arwood, Wehby, & Falk, 2005) where the ratio of treatment to baseline phase *SD*s ranged from 1.2 for Kim to 6.4 for Jack (example data are available in the scdhlm web-based calculator; Pustejovsky, 2016). In some single-case studies the *SD* ratio is more extreme, such as when the baseline values are all zero for a case (e.g., Case 4 in the multiple-baseline study conducted by Laski, Charlop, & Schreibman, 1988). All true values were set to 0.0 for both baseline and intervention phases. Thus, the only difference between phases was in the

variance (i.e., there was no difference in means or slopes), and the differences in variance between phases was equivalent to the differences in the error variances (i.e., the B- to A-phase *SD* ratio was *R*).

As in Investigation 1, the time-series data were generated for multiple-baseline designs with five cases, each with a series length of 22. Again, intervention placement depended on the randomized design that was used. For WW and Rev-M(1) the intervention start points were 6, 9, 12, 15, and 18; for KL(2) and Rev-M(2) the intervention start points were randomly selected from {6, 7}, {9, 10}, {12, 13}, {15, 16}, and {18, 19}; and for MB-R the intervention start points were chosen randomly without replacement from the interval 6 to 18 inclusive.

By crossing the five randomization test methods [WW, KL(2), MB-R, Rev-M(1), and Rev-M(2)] by the effect size (*R* = 1 to 8, in increments of 1) and the level of autocorrelation (ρ = 0, .3), 80 conditions were formed. For each condition, data for 10,000 “studies” were simulated and analyzed. Each data set was analyzed three times: once, using a test statistic operationalized as the average difference in the B- vs. A-phase variances: once, using a test statistic operationalized as the average ratio of the B- to A-phase variances; and once, using a test statistic adapted from Levene’s (1960) and Brown & Forsythe’s (1974) procedures – see Kirk (1995); and Mara & Cribbie (2018) – namely, the difference in the A- and B-phases’ average absolute deviations about their respective phases’ medians. This alternative median-based approach was adopted because in the parametric statistical literature, testing for differences in group variances based on a median-based variance-like statistic has proven to be both more robust and more powerful than testing based on the actual variances themselves (see, for example, Mara & Cribbie, 2018)

Results and Discussion

The results are summarized in Table 4. Paralleling the slope/trend results of Investigation 1, when there are no between-phase differences in variability (i.e., when *R* = 1), for all procedures the Type I error probability was generally maintained at or less than its nominal

level of .05. The one noticeable exception to that statement is with the MB-R procedure applied to the variance-ratio situation, where the empirical αs were .057 and .066 for autocorrelations of 0 and 0.3, respectively.

Concerning power, several aspects of Table 4 are worth noting. First, when the autocorrelation is 0, with few exceptions [e.g., an effect-size of *R* = 2 for the WW and KL(2) procedures] for all five procedures there appears to be a slight power advantage associated with the variance-ratio test relative to the variance-difference test; whereas when the autocorrelation is 0.3 there is a reversal, with the variance-difference test exhibiting slightly but consistently greater power than the variance-ratio test.

Second, as in Investigation 1, power decreases are apparent for all five test procedures with the variance-ratio test as the magnitude of the autocorrelation increases from ρ = 0 to ρ = 0.3. In contrast to that typical pattern, however, with the variance-difference test the magnitude of the autocorrelation has little impact on the resulting powers. For example, for the WW procedure, the comparative powers for ρ = 0 and ρ = 0.3 are .54 vs. .53, .71 vs. .70, and .747 vs. .746 for effect sizes of *R* = 2, 3, and 4, respectively. At the same time, for the KL(2) and MB-R procedures, even slight *increases* in power can be seen as ρ increases from 0 to 0.3. Similar power increases with an increase in the autocorrelation are evident for the KL(2) procedure applied to the median absolute deviation measure. Apart from sampling error, a potential, admittedly after the fact, account of this is that an increase in autocorrelation also increases the total variance (σ_y^2) because $\sigma_y^2 = \frac{\sigma_a^2}{1-\rho^2}$,where σ_a^2 is the variance of the white noise and ρ is the autocorrelation. With an autocorrelation of .3 the total variance is about 1.1 times the variance when the autocorrelation is zero. This change in variance accounts for some of the power decreases when comparing means and slopes because greater variability decreases the standardized mean or slope differences. However, this change in variance will not impact the variance ratio.

Third, for both the variance-difference and variance-ratio tests, there are discernible power differences among the five test procedures. Surprisingly, of the five, the between-case Rev-M(1) procedure is clearly associated with the greatest power, with between-procedure differences sometimes amounting to power increases of .25 to. 30 [see, for example, Rev-M(1) vs. MB-R]. Of the three within-case procedures, the WW procedure is slightly more powerful than the KL(2) and MB-R procedures. Such differences diminish with the variance-ratio test.

All that said, however, the most striking finding of the present investigation is that the newly developed median absolute deviation measure is associated with the greatest power, relative to the variance-difference and variance-ratio measures, for all five test procedures. Those power differences are hardly "trivial," insofar as they sometimes amount to .30 or more. On the median absolute deviation measure, the within-case WW, KL(2), and MB-R procedures are associated with the highest powers, whereas the between-case Rev-M(2) procedure lags far behind.

Finally, at first glance the powers for all test procedures in Table 4 appear to be reasonable, in that – even on the found-to-be-inferior variance-difference and variance-ratio measures – they generally approach or exceed a respectable level of .70 when the effect size *(R*) is greater than 2 or 3. It must be remembered, however, that an *R* of 3 represents one phase's standard deviation being three times larger than the other phase's standard deviation – or equivalently, that one variance is 9 times greater than the other variance. Similarly, an *R* of 4 represents one phase's variance being 16 times larger than the other phase's variance. For readers reared on traditional parametric statistics, these variance differences may seem shockingly large insofar as with moderate sample sizes and a Type I error probability of .05, variance ratios of only 2 or 2-1/2 generally result in statistically significant variance heterogeneity, leading to the conclusion that the two variances differ from one another. Yet, in the present context, when *R* = 2 (equivalent to a variance ratio of 4), unfortunately there is generally only about a 60% chance of detecting a difference between the two variances (see

Table 4). Given these considerations, we regard the present single-case randomization tests of between-phase variability differences based on a series length of 22 and 5 cases to be underpowered (but see also Footnote 2). For that reason, we conducted a supplementary investigation of the median absolute deviation measure with two of the variability test procedures.

Investigation 2a: Supplementary Examination of Two Variability Randomization Tests with the Median Absolute Deviation Measure

Comparable to our study of slopes, we extended the study of between-phase variability differences also based on 5 cases, to examine the impact of series length on the mean difference of the median absolute deviations, which emerged as the most sensitive in our primary Investigation 2 study tests. WW and Rev-M(1) were selected as the test procedures because they were the most powerful of the three within-cases and two between-cases test procedures, respectively, with the median absolute deviation statistic in that investigation (see Table 4). When series length was extended to 32, the potential intervention start points were 7, 12, 17, 22, and 27. When series lengths were extended to 44, the potential intervention start points were 12, 18, 24, 30, and 36. All other methods and conditions were the same as those in the primary study of variability differences.

For the most part, the results from this investigation mimic the patterns based on the 22-observations series length of Investigation 2, but with greater powers. With an autocorrelation of .3 both the WW and Rev-M(1) procedures yielded reasonable powers (.76 and .73, respectively) to detect an effect size of $R = 2$ with a series length of 32; and with a series length of 44, the respective powers were .85 and .83. Thus, the modest power differences favoring WW over Rev-M(1) in Table 4 for a series length of 22 observations at effect sizes of $R = 2$ and 3 disappeared at the longer series lengths of 32 and 44.

As a footnote to these supplementary investigations, comparable power increases might be expected at a series length of 22 by increasing the number of cases from 5 to 6.

Conclusions and Recommendations

In the present single-case multiple-baseline intervention simulation study, rather than the typical focus on between-phase changes in levels (means), here we concentrated on between-phase changes in trends (slopes) and variability. Throughout the manuscript, we discussed the important findings that emerged from our investigation and so, given those along with the Abstract, there is no need to reiterate the specific findings here. Instead in this final section, following a discussion of limitations we provide a few selected conclusions, extensions, and recommendations for practice.

Limitations

As with other simulation studies, the conclusions are limited by the conditions that we examined here. Because this was the first study of the power and Type I error control for randomization tests when they were used to detect slope or variability changes, we focused on conditions where there was just a change in slope, with no immediate change in level or variability, and then on conditions where there was just a change in variability, with no change in level or slope. However, future research should extend the simulations to consider contexts where there are multiple types of change, such as an immediate change in level coupled with an immediate change in slope, or a change in slope coupled with a change in variability – along the lines of the recent work of Tanious et al. (2019). In addition, although we generated data based on normally distributed continuous outcomes, we recognize that single-case intervention studies are often based on count outcomes, which have different base distributions. If future research were to extend the power estimates to count outcomes, as well as to consider additional series lengths and different numbers of cases, a more complete picture would emerge of the power of randomization tests for detecting between-phase slope and variability changes.

Summary of Different Multiple-Baseline Randomization-Test Procedures

Table 5 presents a summary of multiple-baseline randomization-test procedures that have been investigated empirically over the past 25 years or so (starting from Ferron & Ware,

1995), which have examined the statistical properties of tests for assessing level, trend, and variability measures. For each measure, we have rank-ordered the different procedures with respect to their likelihood of detecting intervention effects (i.e., statistical power), given that each of the procedures listed possesses acceptable Type I error control. In these comparisons, *N* represents the number of test procedures that supplied power evidence, and the different procedures were ordered from 1 to *N,* with the method of midranks incorporated for procedures associated with equivalent powers. For each situation considered, we assume that a stable baseline with no baseline trend is present (see, for example, Chen, Yu, & Peng, 2019). The tabled recommendations must be considered with respect to the specific case numbers, series lengths, between-case stagger positions, and number of potential intervention start points on which the simulation studies are based. Not included in the table are data from a study currently in progress that compares the statistical properties of the original Marascuilo-Busk procedure adapted to a multiple-baseline design (i.e., a design formulated to include small staggered intervention start-point intervals for the individual cases) with the Koehler-Levin procedure. Similarly, and as was alluded to earlier, another study in progress is examining the power benefits and tradeoffs between increasing the series lengths and increasing the number of cases.

The first three columns of Table 5 focus on randomization-test procedures that assess between-phase level change in multiple-baseline designs. Based on a review of the various simulation studies in the literature, our recommendations follow. We note again that the Koehler-Levin procedure with three potential intervention start points per case was not included in the present study. As a cautionary comment, in our simulations, series length, number of cases, and amount of delay were held constant, but the amount of between-case stagger was not.

Immediate abrupt effects. Of the various procedures that have been examined, the Koehler-Levin procedure based on two or three potential intervention start points per case and

the restricted Marascuilo-Busk procedure are generally the most powerful and the two modified Revusky procedures are the least powerful (Levin et al., 2018).

Delayed abrupt effects. For delayed abrupt effects, the restricted Marascuilo-Busk procedure is the most powerful, followed by the Koehler-Levin procedure based on two or three potential intervention start points per case. The Wampold-Worsham procedure (equivalent to the Koehler-Levin procedure based on one potential intervention start point per case) is the least powerful of the six procedures listed. It is important to note that the powers of all procedures are considerably lower for delayed abrupt effects than they are for immediate abrupt effects. However, if the amount of delay is correctly anticipated and built into the analysis (as can be effected in Gafurov and Levin's (2020) *ExPRT* single-case randomization test package discussed below) then: (1) the powers improve dramatically to the approximate levels of Immediate abrupt effects; and (2) the comparative powers among the different test procedures remain similar to those for the immediate abrupt effect results (Levin et al., 2017a).

Immediate gradual effects. Here, there is little to choose among the four viable within-series test procedures, but the powers are low and unacceptable for all of them. If an immediate gradual effect is correctly anticipated and built into the analysis (currently not implemented in Gafurov and Levin's, 2020, *ExPRT* package), then: (2) the power levels improve slightly for all test procedures; but (2) the ordering of their comparative powers remains similar to that of the original test procedures' ordering (Levin et al., 2017a).

Assessing Trend (Slope) Between-Phase Change

For between-phase changes in slope investigated in the present study, first we note that because of economies of time and effort, the Koehler-Levin procedure based on three potential intervention start points per case was not included. These results are based on an effect size given by the simple difference in slopes, β_B-β_A. If the change in slope continues throughout the B phase with no asymptote, the Wampold-Worsham procedure is the most powerful and the modified Revusky procedure with two potential intervention start points per case is the least

powerful. With a B-phase asymptote present, there is little difference in the powers of the three within-series procedures (Wampold-Worsham, Koehler-Levin with two potential intervention start points per case, and restricted Marascuilo-Busk).

As a relevant aside, in single-case intervention research phase differences in slope will typically be accompanied by phase differences in levels – as for example, the pattern represented in the upper portion of Figure 1. Such was the situation in the present simulation study of slope differences. In practice, a one- or two-observation delayed change in slope is likely to be indistinguishable from a one- or two-observation delayed gradual change in level in that both would exhibit a gradually increasing slope across sessions. We offer a two-stage recommendation to disambiguate the two. In Stage 1 a test of slope change would be conducted, as in the present study. Then, if that test is statistically significant, in Stage 2 a test of levels would be conducted to infer whether the Stage 1 effect was attributable to a phase change in slope only (i.e., a statistically significant result in Stage 1 but not in Stage 2) or to a phase change in both slope and level (i.e., a statistically significant result in Stage 1 and Stage 2). In addition, for tests of between-phase slope changes, one could compare the powers (at different effect sizes) of one- and two-observation delayed abrupt slope effects with one- and two-observation delayed gradual level effects. Both tests could be conducted (with controlled Type I error probabilities) through the *ExPRT* randomization-test package discussed in the next section. The statistical properties of this proposed two-stage procedure have yet to be empirically validated, however (see also Footnote 2).

Assessing Variability Between-Phase Change

Tests based on the variance-difference measure turned out be somewhat more powerful than those based on the variance-ratio measure for smaller effect sizes. The biggest surprise is that on both the variance-difference and variance-ratio measures the modified Revusky procedure with one potential start point per case proved to overpower all the other procedures. There was little difference between the two within-series procedures [(Wampold-Worsham and

Koehler-Levin(2)] for larger effect sizes. That said, our single-case randomization-test operationalization of the Brown-Forsythe adaptation of Levene's (1960) median absolute deviation test of variances proved to be the clear measure of choice for assessing between-phase changes in variability. Either the within-case Wampold-Worsham procedure, the Koehler-Levin procedure with two potential intervention start points per case, or the restricted Marascuilo-Busk procedure is to be recommended for conducting the randomization test with the median absolute deviation measure.

Available Computer Software

Finally, from a single-case educational intervention researcher's data-analysis standpoint, Gafurov and Levin's (2020) freely available Excel-based single-case randomization test software package (*ExPRT*) initially included the variance-difference and variance-ratio measures to test for A-to-B phase changes in variability. Because of the present simulation's discovery of the superiority of the median absolute deviation measure, that measure is now being used as *ExPRT's* sole basis for statistically assessing between-phase variability changes. Alternatively, for single-case educational intervention researchers who use the statistical software R, Bulté & Onghena's (2013) SCRT package is available for conducting randomization tests for reversal, alternating treatment, and multiple-baseline designs. For tests that are sensitive to slope or variability changes, as we have focused on here, the user would need to customize the test statistic using variable identifiers. The user-friendly website, https://tamalkd.shinyapps.io/scda/, mentioned by Tanious & Onghena (2019), also merits consideration. We hope that educational intervention researchers will take advantage of the single-case designs, analyses, and resources that were cited in this article.

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 communication needs. *Research in Developmental Disabilities*, *56*, 165-176.

Barton-Arwood, S. M., Wehby, J. H., & Falk, K. B. (2005). Reading instruction for elementary-age students with emotional and behavioral disorders: Academic and behavioral outcomes. *Exceptional Children*, *72*, 7-27.

Beretvas, S. N., & Chung, H. (2008). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2:3*, 120-128.

Bouwmeester, S., & Jongerling, J. (2020). Power of a randomization test in a single case multiple baseline AB design. PLOS ONE, 15(2), e0228355. https://doi.org/10.1371/journal.pone.0228355

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*, 364-367.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods*, *12*, 450-478.

Busse, R. T., McGill, R. J., & Kennedy, K. S. (2015). Methods for assessing single-case school-based intervention outcomes. *Contemporary School Psychology*, *19*, 136-144.

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). *Computing tools for implementing standards for single-case designs. Behavior Modification*, *39*, 835-869.

Chen, L.-T., Wu, P.-J., & Peng, C.-Y. J. (2019). Accounting for baseline trends in intervention studies: Methods, effect sizes, and software. *Cogent Psychology*, *6*; retrievable from https://doi.org/10.1080/23311908.2019.1679941.

Collier-Meek, M. A., Sanetti, L. M. H., Levin, J. R., Kratochwill, T. R., & Boyle, A. M. (2019). Evaluating implementation supports delivered within problem-solving consultation. *Journal of School Psychology*, *72*, 91-111.

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

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

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., Joo, S.H., & Levin, J. R. (2017). A Monte-Carlo evaluation of masked-visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis*, *50*, 701-716.

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

Ferron, J., & 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. *ExPRT* (*Excel Package of Randomization Tests): Statistical Analyses of Single-Case Intervention Data*; current Version 4.1 (March 2020) is retrievable from the *ExPRT* website at http://ex-prt.weebly.com

Gast, D. L. (Ed.). (2010). *Single subject research methodology in behavioral sciences*. New York, NY: Routledge.

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

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

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

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & 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. (2019). Application of a single-case intervention procedure to assess the replicability of a two-component instructional strategy. *Contemporary Educational Psychology*, *56*, 161-170.

Hwang, Y., Levin, J. R., & Johnson, E. W. (2018). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design**.** *Developmental Neurorehabilitation*, *21*, 223-237.

Jacobs, K. W. (2019). Replicability and randomization test logic in behavior analysis. *Journal of the Experimental Analysis of Behavior*, *111*, 329-341.

Jamshidi, L., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., & Van den Noortgate, W. (2019, September 1). Bias adjustment in multilevel meta-analysis of standardized single-case experimental data. *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2019.1658568

Joo, S. H., Ferron, J. M., Moeyaert, M., Beretvas, S. N., & Van den Noortgate, W. (2019). Approaches for specifying the level-1 error structure when synthesizing single-case data. *Journal of Experimental Education*, *87*, 55-74.

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

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

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

Kratochwill, T. R., & 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.

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, *21*, 391-400.

Levene, H. (1960). Robust tests of equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Palo Alto, CA: Stanford University Press.

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

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

Levin, J. R., Ferron, J. M., & 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. (2017a). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology*, *63*, 13-34.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017b). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*; *21*. 290-311.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2019). An improved two independent-samples randomization test for single-case AB-type intervention designs: A 20-year journey. *Journal of Modern Applied Statistical Methods*, *18*(1), Article 23, 1-20. Online version available at DOI:10.22237/jmasm/15566/70480.

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., Kratochwill, T. R., & Ferron, J. M. (2019). Randomization procedures in single-case intervention research contexts: (Some of) “The rest of the story”. *Journal of the Experimental Analysis of Behavior*, *112*, 334-348.

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

Maggin, D. M., Cook, B. G., & Cook, L. (2018). Using single-case research designs to examine the effects of interventions in special education. *Learning Disabilities Research & Practice*, 33, 182-191.

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

Manolov, R. (2019). A simulation study on two analytical techniques for alternating treatments designs. *Behavior Modification*, *43*, 544-563.

Manolov, R., & Moeyaert, M. (2017). How can single-case data be analyzed? Software resources, tutorial, and reflections on analysis. *Behavior Modification*, *41*, 179-228.

Manolov, R., & Solanas, A. (2009). Problems of the randomization test for AB designs. *Psicológica*, *30*, 137-154.

Mara, C. A., & Cribbie, R. A. (2018). Equivalence of population variances: Synchronizing the objective and analysis. *Journal of Experimental Education*, *86*, 442-457.

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

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

Michiels, B., Heyvaert, M., & Onghena, P. (2018). The conditional power of randomization tests for single-case effect sizes in designs with randomized treatment order: A Monte Carlo simulation study. *Behavior Research Methods*, *50*, 557-575.

Michiels, B., & Onghena, P. (2019). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*, 2454-2476.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, T., & Van den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719-748.

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.

Parker, R. I., Vannest, K. J., & Davis J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 127-151). Washington, DC: American Psychological Association.

Plavnick, J. B., & Ferreri, S. J. (2013). Single-case experimental designs in educational research: A methodology for causal analyses in teaching and learning. *Educational Psychology Review*, *25*, 549-569.

Pustejovsky, James E. (2016). scdhlm: A web-based calculator for between-case standardized mean differences (Version 0.3.1) [Web application]. Retrieved from: https://jepusto.shinyapps.io/scdhlm

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

Rindskopf, D. M., & Ferron, J. M. (2014). Using multilevel models to analyze singe-case design data (pp. 221-246). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

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

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2014). Analyzing single-case designs: *d*, *G*, hierarchical models, Bayesian estimators, and the hopes and fears of researchers about analyses (pp. 247-281). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

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.

Solanas, A., Manolov, R. & Onghena, P. (2010). Estimating slope and level change in *N* = 1 designs. *Behavior Modification, 34*, 195-218.

Tanious, R., De, T. K., & Onghena, P. (2019). A multiple randomization testing procedure for level, trend, variability, overlap, immediacy, and consistency in single-case phase designs. *Behaviour Research and Therapy*, *119*; retrievable from https://doi.org/10.1016/j.brat.2019.103414.

Tanious, R., & Onghena, P. (2019). Randomized single-case experimental designs in healthcare research: What, why, and how? *Healthcare*, *7*, 143; doi:10.3390/healthcare7040143.

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

Footnotes

The first two authors contributed equally to this study. Correspondence concerning the article should be addressed to Joel R. Levin at jrlevin@u.arizona.edu. We are grateful to two anonymous reviewers, whose revision suggestions greatly improved the quality of our initial submission.

1. In addition, Ferron and Jones (2006) and Ferron, Joo, and Levin (2017) provide discussion a novel response-guided intervention start-point determination strategy that encompasses a solid statistical component.
2. For a sequential procedure to estimate an A- to B-phase change in trend and level, controlling for baseline trend, see Solanas, Manolov, & Onghena (2010). For an alternative randomization-test approach to assessing multiple between-phase change characteristics in ABAB designs, see Tanious et al. (2019).

Table 1. Investigation 1: Type I error (*d* = 0.0) and power for the Wampold-Worsham (WW), Koehler-Levin (KL), revised Marascuilo-Busk (MB-R), and modified Revusky (Rev-M) procedures for detecting changes in trend when there are 5 Cases and 22 observations per case, with autocorrelations (ρ) of 0 and 0.3.

ρ	*d*	$\beta_B - \beta_A$				
		WW	KL(2)	MB-R	Rev-M(1)	Rev-M(2)
0.0	0.0	.049	.033	.032	.047	.047
	0.5	.101	.065	.071	.075	.064
	1.0	.181	.123	.142	.115	.083
	1.5	.281	.200	.239	.151	.109
	2.0	.418	.290	.357	.204	.123
	2.5	.544	.407	.495	.261	.154
	3.0	.663	.508	.638	.342	.191
	3.5	.781	.617	.744	.416	.223
	4.0	.853	.715	.833	.489	.256
0.3	0.0	.046	.039	.034	.052	.054
	0.5	.091	.063	.064	.075	.066
	1.0	.142	.113	.117	.108	.089
	1.5	.232	.170	.187	.142	.108
	2.0	.327	.255	.275	.191	.138
	2.5	.427	.341	.386	.254	.167
	3.0	.548	.444	.505	.312	.201
	3.5	.656	.549	.604	.389	.242
	4.0	.746	.647	.705	.470	.289

Note: The measures are based on the average difference in B and A phase slopes ($\beta_B - \beta_A$). The slope is 0 during baseline and .2*d* throughout the treatment phase, where *d* is the difference between the xpected treatment phase value and the expected baseline value 5 observations into the treatment phase.

Table 2. Investigation 1: Type I error (*d* = 0.0) and power for the Wampold-Worsham (WW), Koehler-Levin (KL), revised Marascuilo-Busk (MB-R), and modified Revusky (Rev-M) procedures for detecting changes in trend when the treatment phase trend asymptotes and there are 5 Cases and 22 observations per case, with autocorrelations (ρ) of 0 and 0.3.

ρ	*d*	$\beta_B - \beta_A$				
		WW	KL(2)	MB-R	Rev-M(1)	Rev-M(2)
0.0	0.0	.055	.033	.033	.052	.047
	0.5	.102	.069	.081	.074	.063
	1.0	.187	.134	.160	.108	.085
	1.5	.276	.217	.255	.149	.100
	2.0	.402	.327	.370	.207	.132
	2.5	.518	.443	.489	.263	.160
	3.0	.622	.545	.583	.324	.181
	3.5	.713	.644	.680	.409	.213
	4.0	.787	.727	.739	.482	.253
0.3	0.0	.052	.036	.035	.047	.048
	0.5	.092	.067	.070	.074	.061
	1.0	.156	.120	.131	.101	.080
	1.5	.229	.196	.200	.143	.105
	2.0	.316	.292	.298	.189	.137
	2.5	.427	.380	.391	.251	.164
	3.0	.515	.468	.497	.321	.202
	3.5	.605	.574	.584	.384	.238
	4.0	.705	.666	.651	.458	.281

Note: The measures are based on the average difference in B and A phase slopes ($\beta_B - \beta_A$). The slope is 0 during baseline, .2*d* throughout the first 4 treatment observations, asymptotes at Observation 5 and throughout the remainder of the treatment observations, where *d* is the difference between the expected treatment phase value and the expected baseline value 5 observations into the treatment phase.

Table 3. Investigation 1 supplementary examination: Type I error (*d* =0.0) and power for the 5-case Wampold-Worsham (WW) and Koehler-Levin [KL(2)] procedures for testing changes in trend as a function of series length (L) , with autocorrelations (ρ) of 0 and 0.3.

		β_B-β_A (no asymptote)						β_B-β_A (asymptote)					
ρ	*d*	L=22		L=32		L=44		L=22		L=32		L=44	
		WW	KL(2)	WW	KL(2)	WW	KL(2)	WW	KL(2)	WW	KL(2)	WW	KL(2)
0.0	0.0	.049	.033	.047	.050	.051	.048	.055	.033	.051	.053	.051	.047
	0.5	.101	.065	.145	.130	.267	.247	.102	.069	.010	.108	.112	.108
	1.0	.181	.123	.313	.280	.626	.572	.187	.134	.177	.181	.192	.181
	1.5	.281	.200	.518	.460	.881	.843	.276	.217	.286	.277	.275	.283
	2.0	.418	.290	.704	.653	.974	.961	.402	.327	.391	.371	.374	.366
	2.5	.544	.407	.853	.791	.999	.991	.518	.443	.494	.475	.451	.456
	3.0	.663	.508	.933	.900	.996	.998	.622	.545	.582	.576	.532	.523
	3.5	.781	.617	.974	.950	1.00	1.00	.713	.644	.662	.669	.585	.598
	4.0	.853	.715	.991	.983	1.00	1.00	.787	.727	.731	.747	.638	.657
0.3	0.0	.046	.039	.047	.049	.052	.050	.052	.036	.050	.052	.052	.049
	0.5	.091	.063	.118	.113	.190	.182	.092	.067	.092	.093	.096	.096
	1.0	.142	.113	.241	.223	.454	.423	.156	.120	.153	.152	.147	.150
	1.5	.232	.170	.393	.370	.722	.682	.229	.196	.224	.216	.211	.216
	2.0	.327	.255	.558	.533	.887	.861	.316	.292	.304	.306	.281	.296
	2.5	.427	.341	.712	.682	.967	.952	.427	.380	.390	.393	.352	.365
	3.0	.548	.444	.838	.806	.991	.989	.515	.468	.474	.483	.415	.427
	3.5	.656	.549	.915	.894	.998	.997	.605	.574	.548	.570	.480	.492
	4.0	.746	.647	.959	.941	1.00	1.00	.705	.666	.632	.641	.527	.544

Note: The table includes powers associated with both the change in trend (no asymptote) and the change in trend that asymptotes 5 observations into treatment (asymptote). *d* is the difference between the expected B-phase value and the expected A-phase value 5 observations into the intervention phase, and the measures are based on the average difference in B- and A- phase slopes (β_B-β_A). For ease of comparison, the L = 22 column power values are copied from Investigation 1.

Table 4. Investigation 2: Type I error (*R* = 1.0) and power for the Wampold-Worsham (WW), Koehler-Levin [KL(2)], restricted Marascuilo-Busk (MB-R), and modified Revusky [Rev-M(1) and Rev-M(2)] procedures for detecting changes in variability according to three different measures, with 5 cases and 22 observations per case, with autocorrelations (ρ) of 0 and 0.3.

ρ	Effect (*R*)	Variance Difference					Variance Ratio					Median Absolute Deviation				
		WW	KL(2)	MB-R	Rev-M(1)	Rev-M(2)	WW	KL(2)	MB-R	Rev-M(1)	Rev-M(2)	WW	KL(2)	MB-R	Rev-M(1)	Rev-M(2)
0.0	1.0	.050	.049	.038	.048	.052	.053	.053	.057	.051	.049	.047	.049	.051	.053	.052
	2.0	.544	.510	.444	.591	.452	.494	.487	.465	.481	.397	.642	.621	.645	.536	.447
	3.0	.708	.670	.608	.857	.664	.709	.713	.677	.764	.662	.890	.876	.903	.830	.704
	4.0	.747	.709	.652	.952	.727	.777	.786	.755	.899	.776	.951	.940	.961	.929	.815
	5.0	.771	.735	.683	.977	.763	.804	.814	.790	.950	.833	.975	.965	.979	.971	.868
	6.0	.783	.744	.683	.987	.782	.817	.826	.801	.972	.872	.978	.976	.986	.985	.889
	7.0	.788	.751	.690	.994	.777	.821	.839	.808	.983	.886	.988	.980	.990	.994	.912
	8.0	.794	.756	.700	.997	.789	.830	.837	.817	.990	.897	.988	.984	.991	.996	.922
0.3	1.0	.052	.050	.052	.049	.052	.051	.049	.066	.050	.050	.049	.045	.050	.050	.051
	2.0	.526	.518	.491	.574	.448	.433	.439	.436	.466	.402	.609	.607	.587	.521	.440
	3.0	.698	.683	.660	.859	.649	.648	.650	.634	.769	.640	.862	.844	.857	.825	.697
	4.0	.746	.736	.710	.950	.719	.717	.725	.701	.898	.755	.928	.917	.934	.932	.799
	5.0	.770	.748	.728	.977	.748	.736	.763	.730	.944	.809	.953	.947	.960	.969	.852
	6.0	.776	.759	.733	.988	.764	.764	.784	.750	.970	.846	.966	.964	.969	.985	.877
	7.0	.782	.763	.745	.993	.773	.766	.784	.763	.983	.858	.975	.965	.977	.992	.896
	8.0	.794	.770	.744	.995	.778	.778	.795	.771	.988	.873	.977	.972	.980	.996	.906

Note: Effect (*R*) is the ratio of the B to A phase standard deviation parameters. The measures were based on either the average difference in B and A phase variances (Difference), the average ratio of the B and A phase variances (Ratio), or the mean difference in the median absolute deviations.

Table 5. Summary table of comparative power ranks, 1 to *N* (1 = Generally most powerful, *N* = Generally least powerful) for the Wampold-Worsham (WW), Koehler-Levin (KL), restricted Marascuilo-Busk (MB-R), and modified Revusky (Rev-M) procedures

	Level			Trend (β_B-β_A)		Variability (Mdn. Abs. Dev.)[c]
	Immediate Abrupt[a]	Delayed Abrupt[b]	Immediate Gradual[b]	No Asymptote[c]	Asymptote[c]	
WW[d]	4	6	***2.5***	***1***	***2***	***2***
KL(2)	***2***	2.5	***2.5***	2.5	***2***	***2***
KL(3)	***2***	2.5	***2.5***	--	--	--
MB-R	***2***	***1***	***2.5***	2.**5**	***2***	***2***
Rev-M(1)[d]	5.5	4.5	5	4	4	4
Rev-M(2)	5.5	4.5	6	5	5	5

Note:

The "best" procedure(s) in each column is/are highlighted in bold italics.

[a] Based on the results of Levin et al. (2018)

[b] Based on the results of Levin et al. (2017)

[c] Based on the results of the present study with 5 cases

[d] For $N > 3$; power = 0 for $N < 4$

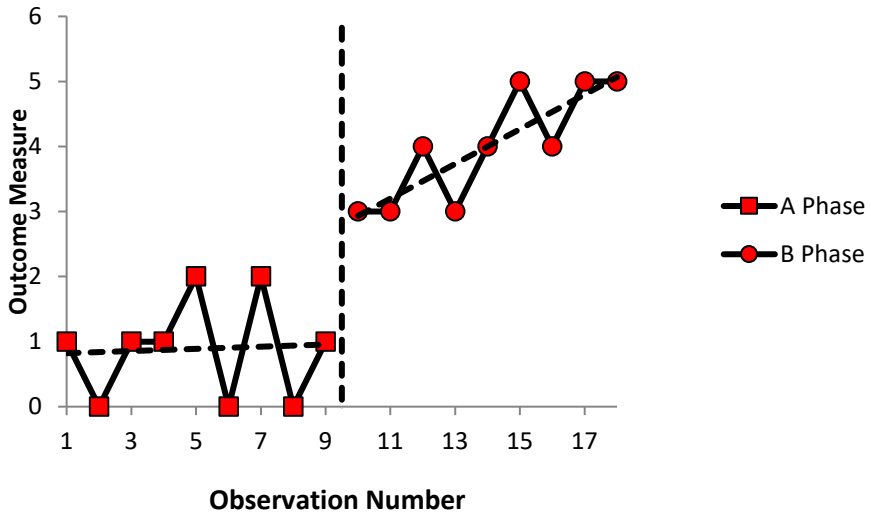
Figure Captions

Figure 1. Hypothetical example of an A- to B-phase increase in slope (Panel 1) and an A- to B-phase decrease in variability (Panel 2)

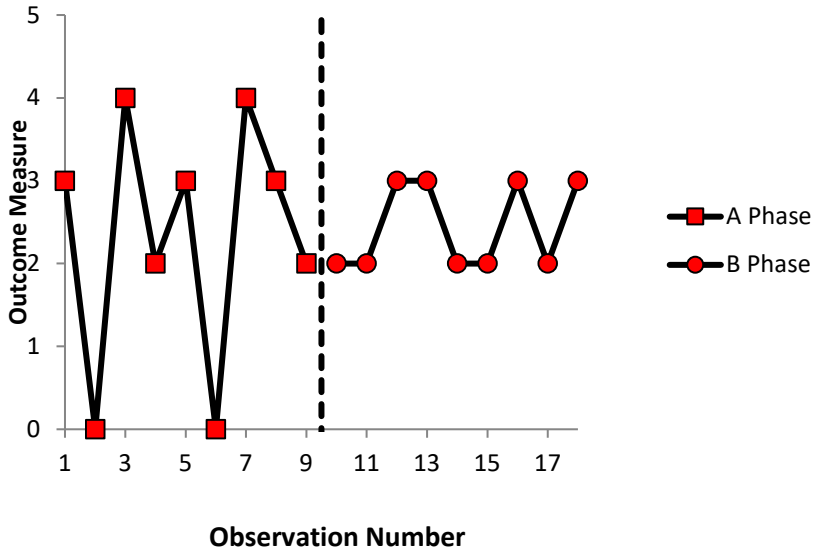
Figure 2. Illustration of prototypical immediate, delayed, abrupt, and gradual between-phase changes in level

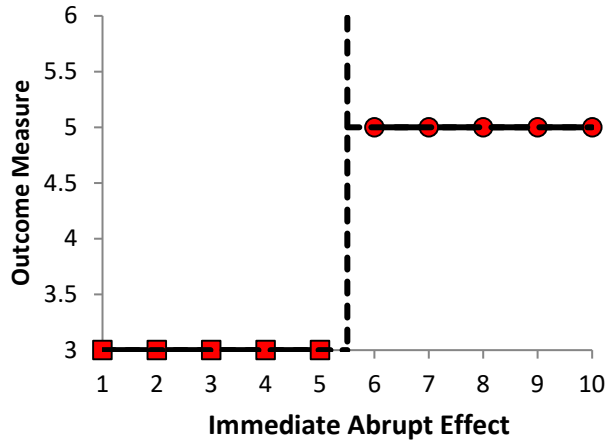
Figure 3. Depiction of effect size (d and b_B - b_A) and associated empirical powers for the Wampold-Worsham test based on 22 observations for each of 5 cases, with r = .30, α = .05, and no asymptote in the intervention slope.

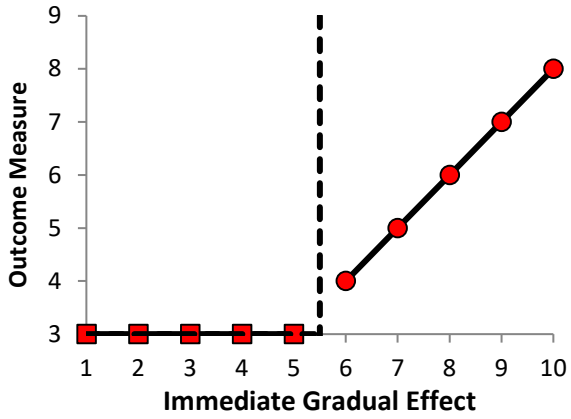
Panel 1

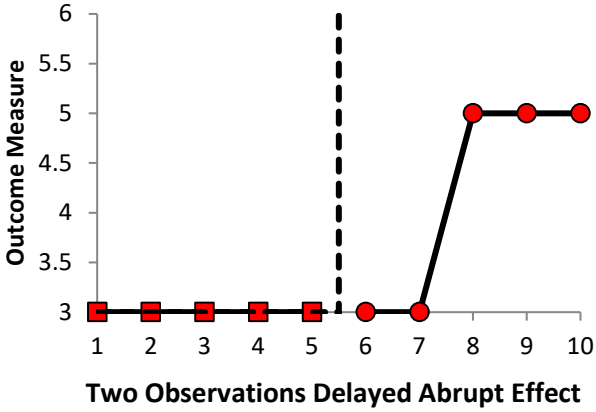


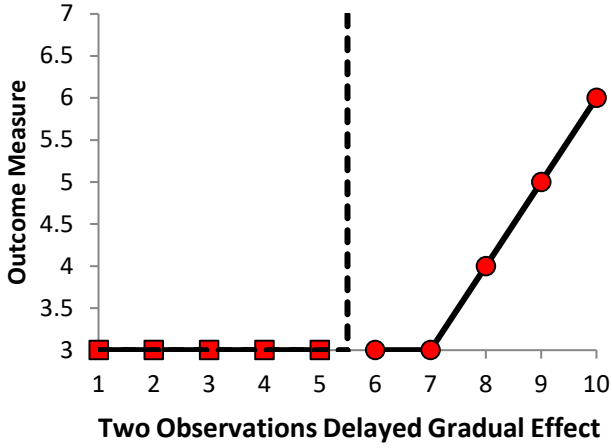
Panel 2

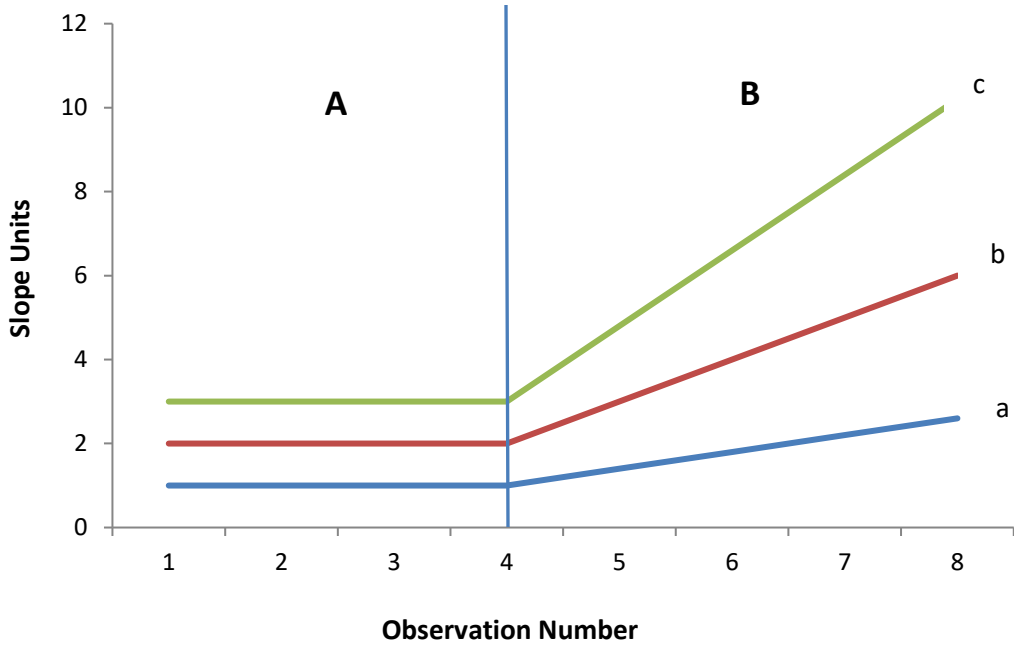



Outcome Measure
6
5.5
5
4.5
4
3.5
3
1 2 3 4 5 6 7 8 9 10
Immediate Abrupt Effect


Outcome Measure
9
8
7
6
5
4
3
1 2 3 4 5 6 7 8 9 10
Immediate Gradual Effect


Outcome Measure
6
5.5
5
4.5
4
3.5
3
1 2 3 4 5 6 7 8 9 10
Two Observations Delayed Abrupt Effect


Outcome Measure
7
6.5
6
5.5
5
4.5
4
3.5
3
1 2 3 4 5 6 7 8 9 10
Two Observations Delayed Gradual Effect



Notes:

[a]***d*** = 2.0, b_B-b_A = 0.4, Power = .33

[b]*d* = 3.0, b_B-b_A = 0.6, Power = .55

[c]*d* = 4.0, b_B-b_A = 0.8, Power = .75