Evaluating Statistical Methods for Single-Case Designs Under Serial Dependence: Precision, Validity and Sensitivity

Hsiu-Ting Yu [1,2]

[1] Department of Psychology, National Chengchi University, Taipei, Taiwan
[2] Research Center for Mind, Brain, and Learning, National Chengchi University, Taipei, Taiwan

©American Psychological Association, 2026. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: https://doi.org/10.1037/met0000879

Author Note

Hsiu-Ting Yu https://orcid.org/0000-0002-3668-8033
This study used simulated data only and did not involve human participants; therefore, ethical approval was not required.

The author was solely responsible for the conceptualization, design, analysis, and writing of this manuscript.

In accordance with APA Transparency and Openness guidelines, the code behind the simulation and analyses has been made publicly available at the Open Science Framework (OSF) repository: https://doi.org/10.17605/OSF.IO/SEZ2T

There is no known conflict of interest to disclose.

This research originated from a project funded by the National Science and Technology Council of Taiwan (NSTC 111-2410-H-004-168) during 2022–2023. Preliminary work related to this project was presented at the International Meeting of the Psychometric Society in 2023. The simulation design, the set of methods compared, and the analyses reported in the present manuscript differ substantially from that presentation. The data and ideas appearing here have not otherwise been disseminated, posted to a website, or shared on a listserv.

Correspondence concerning this article should be addressed to Hsiu-Ting Yu, Department of Psychology, National Chengchi University, No. 64, Section 2, Zhinan Road, Wenshan District, Taipei City 11605, Taiwan. Email: hsiutingyu@gmail.com / htyu@nccu.edu.tw

Abstract

Single-case experimental designs (SCEDs) involve short, serially dependent time series that complicate statistical inference. This Monte Carlo simulation evaluated six methods (NAP, Tau-U, Hedges' *g*, piecewise regression, randomization tests, and Bayesian generalized linear models) on Type I error, power, and interval coverage, each assessed for the methods that admit it, across series lengths (N = 15, 30, 60), autocorrelation (ρ = 0, 0.3, 0.6), baseline trend, and effect structure within an AB design. The two regression-based methods showed Type I error rates reaching .27 to .37 at $\rho = 0.6$ and coverage as low as .62. NAP Type I error inflated with autocorrelation and reached 1.00 under a baseline trend. Coverage for Hedges' *g* fell from .95 to .69 under a flat baseline and collapsed to .00 under a trending baseline. Because these methods did not control Type I error, their nominally higher power is not commensurable with that of a valid test: at a common calibrated Type I error rate the randomization test was at least as powerful as either regression method in every equal-allocation level-change condition at $N \geq 30$. The apparent validity–sensitivity trade-off is therefore largely an artifact of comparing tests that operate at different actual significance levels, compounded by applying a randomization test to a design containing no randomization. Randomizing the intervention start point restored the test to the largest significance level its discrete reference distribution admits. Type I error control is treated as a prerequisite, and design-level randomization is recommended as the principal remedy.

Keywords: single-case experimental design, autocorrelation, Monte Carlo simulation, randomization tests, coverage probability

Translational Abstract

Researchers in applied behavior analysis, special education, and clinical psychology track how a treatment affects a single individual over time rather than comparing large groups. Such studies record repeated measurements before and after an intervention to determine whether observed change reflects a genuine effect. Method choice matters: different methods can yield different conclusions. Two features make this choice difficult: the measurement series are short, and consecutive observations tend to resemble one another, a property called serial dependence. This study evaluated six statistical methods on three criteria: avoiding false-positive conclusions when no effect is present, detecting genuine effects, and providing well-calibrated uncertainty estimates. Under serial dependence, regression-based and Bayesian approaches produced misleadingly narrow uncertainty estimates and excessive false-positive rates, an illusion of precision. Much of their apparent advantage in detecting effects was an artifact of those false alarms: compared at the same actual false-alarm rate, their advantage for immediate treatment effects largely disappeared. A supplementary experiment showed why randomization tests had seemed insensitive: they were applied to studies in which nothing was randomized. When the intervention start point is randomized, the false-alarm rate matches what the design permits at every level of dependence, with no assumptions about the data. Gradual trend changes remained hardest for every method to detect. Sound inference therefore begins with the design: randomize the start point where circumstances permit, plan for thirty or more observations, use randomization tests for confirmatory decisions, describe the consistency of change with nonoverlap indices, and reserve autocorrelation-aware regression for detailed estimation.

Evaluating Statistical Methods for Single-Case Designs Under Serial Dependence: Precision, Validity and Sensitivity

Single-case experimental designs (SCEDs) constitute the methodological backbone of applied behavior analysis, special education, and clinical psychology, enabling researchers to establish causal inferences regarding intervention effects at the individual level (Horner et al., 2005; Kazdin, 2021). Unlike group-comparison designs, SCEDs involve repeated measurement of a single entity across temporally distinct phases (e.g., baseline vs. intervention). Visual analysis has historically served as the primary evaluation method, with proponents arguing that it detects clinically significant changes while filtering out trivial effects (Parsonson & Baer, 1992). However, research has documented low inter-rater reliability in visual analysis (Danov & Symons, 2008; DeProspero & Cohen, 1979), and its inability to quantify effect sizes for meta-analytic aggregation has precipitated a shift toward supplementary statistical analysis (Ferron et al., 2014; Kratochwill et al., 2010; Shadish et al., 2014).

Quantification of SCED data has thus become an increasingly important complement to visual analysis (Tate et al., 2013). SCED data possess structural characteristics that violate fundamental assumptions of conventional parametric statistics: short time series, frequent baseline trends, and, most critically, serial dependence or autocorrelation (Huitema & McKean, 1998; Solomon, 2014).

The Autocorrelation Controversy

Autocorrelation poses a severe threat to the validity of the general linear model, and the debate over its prevalence has been vigorous: early work by Huitema (1985; see also Huitema & McKean, 1998) suggested that autocorrelation might be negligible, justifying OLS approaches. However, large-scale reviews of published data document non-trivial

autocorrelation in applied settings (Pustejovsky et al., 2023; Shadish & Sullivan, 2011; Solomon, 2014). When positive autocorrelation is present but unmodeled, OLS standard errors are biased downward, creating an illusion of precision that leads to inflated test statistics and substantially elevated Type I error rates (Ferron, 2002; Manolov & Solanas, 2008).

The Landscape of Analytical Choices

Applied researchers face a wide array of analytical choices, ranging from nonoverlap indices such as Non-overlap of All Pairs (NAP; Parker & Vannest, 2009) and its trend-corrected extension Tau-U (Parker et al., 2011), to complex parametric approaches including piecewise regression estimated by ordinary least squares (Center et al., 1985; Huitema & McKean, 2000; Moeyaert et al., 2014) and generalized least squares models that explicitly accommodate autocorrelation (Maggin et al., 2011; Swaminathan et al., 2014b). Computer-intensive techniques, such as randomization tests (Edgington, 1980; Edgington & Onghena, 2007) and Bayesian methods (de Vries & Morey, 2013; Natesan & Hedges, 2017; Rindskopf, 2014; Swaminathan et al., 2014a), offer further alternatives. Previous simulation studies have examined randomization tests under various conditions (Bouwmeester & Jongerling, 2020; Ferron et al., 2003; Ferron & Sentovich, 2002; Levin et al., 2018, 2021; Michiels & Onghena, 2019; Michiels et al., 2020), Bayesian approaches (Baek & Ferron, 2021; de Vries et al., 2015; Joo & Ferron, 2019; Natesan Batley & Hedges, 2021), and regression-based methods (Baek et al., 2020; Chen et al., 2025; Li et al., 2025). However, few studies have directly compared methods from all three paradigms within a single simulation framework.

Gaps and Contributions

Three gaps motivate the present study. First, coverage probability has been examined for individual SCED analytical methods (e.g., Baek et al., 2020; Chen et al., 2025; Li et al.,

2025; Natesan Batley & Hedges, 2021), and the present study follows that line of work; what has not been available is a single framework in which methods from all three analytical paradigms are evaluated under identical conditions, with each criterion assessed for the methods that admit it, so that the criteria can be weighed against one another for the same data. Coverage probability, the likelihood that a confidence or credible interval contains the true parameter, is essential for clinical decision-making. Second, most simulations treat level and slope changes as mutually exclusive; in practice, interventions frequently produce combined effects that create parameter confounding. Third, the performance of randomization tests in extremely short series ($N = 15$), where the randomization space is severely restricted, warrants systematic evaluation alongside other methods, as does the question of what a randomization test is estimating when it is applied, as is common in applied practice, to a design in which nothing was randomized.

The present study addresses these gaps through a large-scale Monte Carlo simulation evaluating six analytical methods across a factorial design varying series length, autocorrelation, baseline trend, and effect structure within an AB phase design. By jointly assessing validity (Type I error), sensitivity (power), and precision (coverage), the study provides an evidence-based decision framework for SCED researchers. Two analytic commitments shape the interpretation. First, control of the Type I error rate is treated as a prerequisite for a test rather than as a quantity that may be traded against power: power comparisons between tests whose actual significance levels differ are not interpretable, because the more liberal test rejects more often under null and alternative alike. Power is therefore reported both at the nominal .05 level and at an empirically calibrated critical value equating actual Type I error rates across methods. Second, the inferential warrant of a randomization test

derives from randomization in the design, not from the computational act of enumerating data divisions (Edgington, 1980). The main simulation deliberately reproduces the prevailing applied practice of a fixed intervention start point; a supplementary simulation (Study 4) contrasts it with the randomized design that the logic of the test requires.

Method

Statistical Approaches Evaluated

Six methods were selected to represent the three dominant paradigms in SCED analysis: non-parametric overlap indices, parametric regression, and computer-intensive inferential techniques.

Non-overlap of All Pairs (NAP)

NAP utilizes all available data pairs, making it more sensitive to distributional shifts and less susceptible to outliers relative to alternative overlap statistics such as the Percentage of Nonoverlapping Data (PND; Scruggs et al., 1987). The NAP index is calculated as:

$$NAP = \frac{\text{Pairs}_{B>A} + 0.5 \times \left(\text{Pairs}_{B=A}\right)}{\text{Total Pairs}}.$$

NAP yields a score between 0 and 1, where 0.50 indicates chance-level overlap. NAP quantifies ordinal dominance, the estimated probability that a randomly selected intervention-phase observation exceeds a randomly selected baseline-phase observation, on a probability metric rather than on the metric of the outcome. It therefore does not express magnitude as a distance, and benchmarks developed for distance-based effect sizes do not transfer to it (Manolov & Tanious, 2025; Pustejovsky, 2019). For this reason the present article refers to NAP and Tau-U as nonoverlap indices rather than as effect-size measures. A critical limitation is that NAP does not adjust for baseline trend: an improving pre-intervention trajectory inflates apparent non-overlap, confounding trend effects with intervention effects. In this sense, NAP is

sensitive to baseline trend but does not control for it. It should be noted that NAP, like other nonoverlap indices, has a ceiling effect that can compromise its interpretation as a continuous index of treatment magnitude (Carter, 2013; Manolov & Tanious, 2025).

Tau-U

Tau-U extends Kendall's rank correlation coefficient by combining the ordinal comparison of phases with a correction for monotonic baseline trend (Parker et al., 2011). Tau-U allows researchers to correct for monotonic trends in Phase A before comparing phases. The calculation involves examining the ordinal relationship between all pairs across phases (S_{AvsB}) and subtracting the trend component within Phase A (S_{TrendA}):

$$\text{Tau-}U = \frac{S_{AvsB} - S_{TrendA}}{n_A \times n_B}.$$

Two clarifications about this expression are required. First, it corresponds to the variant of Tau-U that controls for baseline trend but does not incorporate an intervention-phase trend term; the variants that add a Phase B trend component, or that omit the baseline correction, are not evaluated here. Second, the denominator is not uniquely defined in the source literature. Parker et al. (2011) report in their Table 1 a pair count that adds the $n_A(n_A-1)/2$ within-baseline comparisons to the $n_A \times n_B$ cross-phase comparisons, whereas the worked example on their p. 293, subsequent restatements of the index, and the widely used online calculator all divide by $n_A n_B$ alone (see Pustejovsky et al., 2024). The present simulation implements the latter convention, the one most applied researchers encounter; the former yields smaller absolute values and cannot exceed unity, whereas the convention used here can. This ambiguity is a property of the index as deployed in practice and is one reason its interpretation remains contested (Brossart et al., 2018).

The null value for Tau-U is 0 (not 0.5 as for NAP), reflecting the expectation of no

systematic ordering of data across phases after baseline trend correction. While Tau-U offers a mechanism for correcting monotonic baseline trends, its interpretation has been questioned (Brossart et al., 2018), and Tarlow (2017) has proposed baseline-corrected Tau (Tau*BC*) as an improved alternative. The present study retains Tau-U rather than Tau*BC* because the simulation evaluates methods as currently deployed in applied practice, informing researchers about the consequences of their actual analytic choices.

Hedges' g (corrected for SCED)

Hedges' g aligns SCED effect sizes with those used in group-design meta-analyses (Hedges et al., 2012). Unlike the standard Cohen's d, the SCED-specific g incorporates a small-sample correction factor (J) to mitigate the bias inherent in estimating population variance from short time series:

$$g = d \times J \text{, where } d = \frac{\overline{Y_B} - \overline{Y_A}}{S_{pooled}}.$$

The small-sample correction factor is $J(m) = 1 - 3/(4m - 1)$, where $m = N - 2$ represents the degrees of freedom for within-case estimation; J corrects the positive bias in sample standard deviation estimates from short series. The within-case standardized mean difference used here differs from the design-comparable between-case standardized mean difference (Hedges et al., 2012, 2013), which models autocorrelation and allows direct pooling with group-design studies. The present within-case formulation was selected to evaluate a commonly encountered practice, namely computing standardized mean differences without explicitly modeling autocorrelation. Interpretive benchmarks for this within-case version differ from those of Cohen's d applied to independent groups.

Piecewise regression

Piecewise regression using OLS provides a flexible framework to model both level and slope changes (Center et al., 1985; Moeyaert et al., 2014). The basic structural model is defined as:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 \left(T_t \times D_t \right) + e_t .$$

The ordinary least squares specification evaluated here follows the analytic tradition established by Center et al. (1985) and examined more recently by Moeyaert et al. (2014). Regression-based alternatives that explicitly model autocorrelation via generalized least squares also exist (Maggin et al., 2011; Swaminathan et al., 2014b), and multilevel extensions can accommodate autocorrelation across participants (Baek & Ferron, 2013). The present OLS implementation was selected to evaluate the consequences of ignoring autocorrelation, a common practice.

Randomization tests

A randomization test constructs its reference distribution from the set of assignments that the experimental design could have produced, and its inferential warrant derives from the fact that one of those assignments was actually selected at random (Edgington, 1980; Edgington & Onghena, 2007). Where that random selection is present, the test requires no distributional assumption and is therefore unaffected by serial dependence in the errors. Where it is absent, the enumeration of data divisions is still computable, but it no longer has a design-based justification; this distinction is developed in the Simulation Design section and evaluated empirically in Study 4. The random intervention start-point technique was used: the time-series order is preserved, and all scientifically plausible intervention start points are considered. For level-change analyses the test statistic was the absolute difference between the intervention-phase and baseline-phase means, evaluated two-tailed; for slope-change analyses it was the

absolute difference between the fitted phase slopes (cf. Levin et al., 2017, 2021). A directional statistic would in fact be the logical choice in most applications, not merely a defensible one: the aim of an intervention is ordinarily known in advance to be an increase or a reduction of the target behavior, and only one of those directions is therapeutically desirable, so a one-tailed statistic aligned with the anticipated direction would yield uniformly higher power. The two-tailed absolute statistic was retained as the specification implemented in the archived simulation code and the more conservative choice. Test-statistic consequences, substantial for the slope analyses, are examined in Study 4.

Where randomization is present in the design, randomization tests control the Type I error rate exactly, up to the discreteness of the reference distribution, whatever the dependence structure of the errors. Their power, however, is bounded by the number of admissible assignments: in a severely truncated series the minimum attainable *p*-value can exceed .05, rendering significance at that level unattainable irrespective of the true effect. Both properties are quantified for the present conditions in the Evaluation Criteria section.

Bayesian generalized linear models

Bayesian methods combine the likelihood of the data with prior distributions to yield a posterior distribution of the treatment effect (Rindskopf, 2014; Shadish et al., 2013). The implementation evaluated here is an approximate Bayesian GLM fitted by penalized maximum likelihood with independent weakly informative Cauchy priors on the scaled coefficients, as implemented in the arm package (Gelman et al., 2008). It was selected as the point of entry requiring the least specialized expertise rather than as the most capable option: the model is specified with the same formula syntax as an ordinary GLM and fitted by a single function call, no sampler must be configured, tuned, or monitored, and the priors are the weakly informative

defaults that Gelman et al. (2008) proposed precisely to serve as defaults. Previous simulation evaluations of Bayesian estimation for single-case data have instead examined fully sampled Markov chain Monte Carlo implementations, typically specified in BUGS-family software and equipped with noninformative or vague priors (Baek et al., 2020; Baek & Ferron, 2021; Moeyaert et al., 2017); those implementations require model specification in a dedicated language and convergence monitoring, and their properties therefore speak to a more technically committed mode of use. The approximate implementation examined here characterizes instead what Bayesian machinery delivers at its lowest barrier to entry. Posterior summaries derive from a normal approximation at the posterior mode rather than from Markov chain Monte Carlo sampling, so convergence diagnostics such as R-hat and effective sample size do not apply, and the 95% intervals are normal-approximation credible intervals. The weakly informative priors introduce shrinkage but carry no substantive information about the size of intervention effects; whether that is a defensible use of Bayesian machinery is addressed in the Discussion.

Simulation Design

Data generation

Time-series data were generated using a first-order autoregressive (AR(1)) process combined with linear trend and treatment effect components. The data-generating model, standard in SCED simulations (see also Manolov & Solanas, 2008; Solomon, 2014), was:

$$y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 \left[\left(T_t - T_{start} \right) \times D_t \right] + u_t$$

where β_0 is the baseline intercept, β_1 is the baseline trend, β_2 represents the immediate level change, and β_3 represents the change in slope. D_t is a dummy variable indicating the phase (0

for baseline, 1 for intervention). Crucially, the error term u_t was generated to reflect serial dependence: $u_t = \rho u_{t-1} + \epsilon_t, \epsilon_t \sim N(0, 1)$, where ρ represents the autocorrelation coefficient.

Three distributional assumptions were made. Errors were drawn from a normal distribution, consistent with the majority of prior SCED simulation studies (Manolov & Solanas, 2008; Solomon, 2014); robustness to non-normality is addressed in the Limitations section. Error variance was assumed homogeneous across time, enabling a clean decomposition of autocorrelation effects; variance heterogeneity models (Baek & Ferron, 2013) represent a meaningful extension. Trends were modeled as linear, consistent with the piecewise regression framework; non-linear growth functions are a further extension.

Experimental Conditions

A full factorial design was used, crossing series length (3 levels) × autocorrelation (3 levels) × baseline trend (2 levels) × effect magnitude (3 levels) × phase allocation ratio (2 levels), yielding 108 conditions in each of Studies 1 and 2. Each condition was replicated 1,000 times. The intervention start point was held fixed within each condition, that is, the design was systematic and contained no randomization. Start-point randomization is uncommon in published applied SCED research, yet randomization tests are used for such data (Tanious & Onghena, 2021); the main simulation reproduces that combination to document its consequences, without endorsing it. Applying a randomization test to a design in which nothing was randomized is not a defensible analytic strategy, because there is no assignment mechanism from which the reference distribution can derive its meaning (Edgington, 1980; Ferron et al., 2003). The conditions reported in Studies 1 through 3 should therefore be read as an evaluation of the consequences of a prevailing but unwarranted practice, not as a recommendation. Study 4 supplies the corresponding randomized-design benchmark. The

randomization test reference distribution was approximated via Monte Carlo sampling: 500 admissible start points were drawn with replacement from the set of possible intervention positions for each replication. Because the reference distribution was sampled rather than enumerated, realized *p*-values deviate slightly from exact multiples of 1/*K*; the smallest realized *p*-values at $N = 15$ were .072 in the pure-null condition of Table 1 and .058 over all $N = 15$ conditions, against the exact bound of 1/9 = .111; all exceed .05, leaving the $N = 15$ power conclusions unaffected. Conditions were selected from published surveys of SCED characteristics (Pustejovsky et al., 2023; Shadish & Sullivan, 2011; Tanious & Onghena, 2021) and prior simulation precedent, with justification as follows.

Series Lengths (N)

Series lengths (N) were set to 15, 30, and 60, where N denotes the total number of observations in a single AB comparison. Published surveys report median total series lengths of 16 observations (Shadish & Sullivan, 2011) and 20 observations (Tanious & Onghena, 2021). Measured against those distributions, $N = 15$ is the condition that corresponds to routine applied practice, $N = 30$ represents the better-powered tail of the empirical distribution, and $N = 60$ lies well beyond it. The inclusion of $N = 60$ is not a claim about typical practice, and the point deserves emphasis because it cuts against the design rather than for it: a replicated design containing several AB comparisons within a realistic total series of 40 to 60 observations yields individual phase contrasts closer to $N = 15$ or $N = 20$ than to $N = 60$. $N = 60$ is retained as an asymptotic benchmark separating small-sample artifacts from failures that persist as information accumulates; the parametric Type I error inflation proves to be of the second kind. Conclusions bearing on applied practice are drawn from the $N = 15$ and $N = 30$ conditions.

Serial Dependence (ρ)

Levels of serial dependency (ρ) were set at 0.0, 0.3, and 0.6. Solomon (2014) aggregated school-based single-case datasets and reported a mean absolute lag-1 autocorrelation of approximately .36, and Shadish et al. (2013) estimated lag-1 autocorrelations in single-case data with a median near .2 and substantial variance; recent evidence confirms that moderate-to-high autocorrelation is prevalent in contemporary SCED research (Tanious & Onghena, 2026). The same three values, 0, .3, and .6, were used by Michiels and Onghena (2019) on the same empirical grounds. The value $\rho = 0$ establishes a baseline of independence, $\rho = 0.3$ approximates the central tendency of the empirical distribution, and $\rho = 0.6$ represents a high-dependence scenario that challenges all methods.

Baseline trend (β_1) and Effect structure (β_2, β_3)

The parameters of baseline trend (β_1) and effect structure (β_2, β_3) governed the shape of the outcome trajectory. All three are expressed in units of the innovation standard deviation of the error process, which was fixed at 1. Baseline trend (β_1) was either flat ($\beta_1 = 0$) or upward ($\beta_1 = 0.2$ in the level-change and combined-effect conditions, 0.1 in the slope-change conditions), a positive secular trend from which the intervention effect must be distinguished. Effect structure manipulated the intervention signal through the level-change (β_2) and slope-change (β_3) parameters, in three configurations: a level-change-only condition, in which β_2 varied across three magnitudes (0, 0.5, and 1.0) while β_3 was fixed at zero; a slope-change-only condition, in which β_3 varied across three magnitudes (0, 0.1, and 0.3) while β_2 was fixed at zero; and a combined-effect condition, in which both parameters were simultaneously nonzero ($\beta_2 = 0.5$, $\beta_3 = 0.1$), representing the realistic scenario of an immediate discontinuity combined with a sustained change in trajectory. These magnitudes follow documented precedent: level changes of 0.5 and 1.0 error standard deviations are the values for which Ferron and Sentovich

(2002) reported randomization-test power and lie within the range examined by Levin et al. (2018); a slope change of 0.1 per occasion has represented a small trend effect in Joo et al. (2019) and Levin et al. (2021), with 0.3 extending the range upward, bracketed by the .25 increments of Michiels and Onghena (2019); and the trending-baseline values $\beta_1 = 0.2$ and 0.1 are deliberately smaller than the trend weights of .25 and .50 used by Michiels and Onghena (2019), so they represent conservative secular trends. One consequence should be noted: because the marginal standard deviation of a stationary AR(1) process is $(1 - \rho^2)^{\wedge}(-1/2)$, a fixed β_2 corresponds to a standardized mean difference that declines with autocorrelation (0.50, 0.48, and 0.40 at $\rho = 0$, 0.3, and 0.6 for $\beta_2 = 0.5$). Power comparisons across autocorrelation levels are therefore mildly conservative; comparisons among methods within a level, which carry the substantive argument, are unaffected. Together, these conditions span the principal expressions of a causal effect in an interrupted time series.

Phase allocation ratios

Two phase allocation ratios were included as a simulation factor: equal allocation (ratio = 0.50, i.e., equal-length phases) and imbalanced allocation (ratio = 0.33, i.e., approximately one-third baseline and two-thirds intervention). Exact phase lengths by condition are as follows: for $N = 15$, ratio = 0.50: $n_A = 7$, $n_B = 8$; ratio = 0.33: $n_A = 4$, $n_B = 11$. For $N = 30$, ratio = 0.50: $n_A = 15$, $n_B = 15$; ratio = 0.33: $n_A = 9$, $n_B = 21$. For $N = 60$, ratio = 0.50: $n_A = 30$, $n_B = 30$; ratio = 0.33: $n_A = 19$, $n_B = 41$. Note that equal allocation is not achievable at $N = 15$ because 15 is odd; the 7/8 split is the closest approximation.

The imbalanced ratio reflects a common applied scenario in which baseline phases are kept short to accelerate intervention delivery. The combined simulation (Study 3) used $N = 30$ with ratio = 0.50 ($n_A = 15$, $n_B = 15$). Although the AB design does not meet the What Works

Clearinghouse (2022) design standards requiring at least three demonstrations of an effect, it is the fundamental unit upon which more complex replicated designs are built. The scope of generalization differs sharply by method. For the model-based methods, the documented properties attach to a single phase contrast and carry over qualitatively to the corresponding contrasts within replicated designs, subject to multilevel structure. For the randomization test, no such transfer is available. The reference distribution of a randomization test is defined by the randomization scheme of the specific design, and those schemes differ fundamentally across designs: phase-order randomization in ABAB designs, staggered start points across participants in multiple-baseline designs, and restricted alternation in alternating-treatments designs each generate a different assignment set of a different cardinality, and hence different attainable significance levels and different power (Levin et al., 2018; Onghena, 1992; Onghena & Edgington, 1994, 2005). The randomization-test results reported here apply to the AB start-point scheme only, and no direct extrapolation to other designs is warranted.

Evaluation Criteria

Method performance was evaluated using three criteria. Type I error rate: the proportion of replications yielding a statistically significant result ($p < .05$ for frequentist methods; 95% credible interval excluding zero for the Bayesian GLM) when the true effect was zero. Type I error rates were benchmarked against Bradley's (1978) liberal criterion, which defines an acceptable range of $[.5\alpha, 1.5\alpha] = [.025, .075]$ for $\alpha = .05$. Statistical power: the proportion of replications correctly detecting a true non-zero effect. Coverage probability: the proportion of 95% confidence intervals (frequentist) or credible intervals (Bayesian) containing the true parameter value. A fourth quantity, size-adjusted power, is reported alongside nominal power. Nominal power is not comparable across tests whose actual Type I

error rates differ: a test that rejects .27 of the time under the null also rejects more often under the alternative without being more sensitive. Size-adjusted power was computed by taking, within each design cell and for each method, the 5th percentile of the null-condition *p*-values as an empirically calibrated critical value, and recording the proportion of replications under the alternative at or below it; it estimates the sensitivity each method would have if it operated at an actual significance level of approximately .05. It is a diagnostic, not a procedure; an applied researcher cannot perform the calibration, which requires knowledge of the true autocorrelation.

Coverage probability was evaluated for Hedges' g, piecewise regression, and the Bayesian GLM only. No confidence interval was computed for NAP or Tau-U at any point in the simulation, and consequently no coverage probability, interval width, or interval-based inference is reported for either index anywhere in this article. The reasons differ: the standard error formula available for NAP assumes independent observations and is known to underestimate true sampling variability under autocorrelation, so an interval computed from it would not be interpretable; and the sampling distribution of Tau-U with baseline correction has not been formally described, so no standard error exists to compute (Pustejovsky et al., 2024). Type I error and power for NAP were evaluated separately using Mann–Whitney *p*-values derived from the normal approximation to the Wilcoxon null distribution, acknowledging that this approximation assumes independence and applies no tie or continuity correction; no analogous computation was performed for Tau-U given its undefined sampling distribution.

For the randomization test, the minimum baseline phase length was set at three observations, and the admissible start points were the positions leaving at least three observations before and at least four observations after the phase change. This yields the

following randomization spaces: $N = 15$: 9 admissible start points, minimum exact p-value = 1/9 = .111; $N = 30$: 24 start points, minimum exact p-value = 1/24 = .042; $N = 60$: 54 start points, minimum exact p-value = 1/54 = .019. Accordingly, statistical significance at $\alpha = .05$ is mathematically impossible at $N = 15$ for the enumerated test, and no realized p-value fell below .058 in the sampled approximation, making power exactly 0 in every replication rather than merely negligible. The largest significance level attainable at or below .05 is the largest multiple of $1/K$ not exceeding it: .000, .042, and .037 at $N = 15$, 30, and 60. An exactly valid randomization test therefore appears conservative against a nominal .05 benchmark, and these attainable levels are the correct benchmark for Studies 1 and 2. It is also possible to construct confidence intervals for single-case effect sizes via randomization test inversion (Edgington & Onghena, 2007; Michiels et al., 2017); this approach was not pursued in the present study, which focuses on hypothesis testing as the primary inferential goal for the randomization test.

Transparency and Openness

All data generation, analysis, and visualization were conducted in R (R Core Team, 2024). Simulation code and output are available at the Open Science Framework repository: https://doi.org/10.17605/OSF.IO/SEZ2T. The simulation was not preregistered.

Results

The results are organized by four analyses: the three intervention scenarios of (a) immediate level change, (b) gradual slope change, and (c) combined effects, followed by (d) a supplementary comparison of systematic and randomized assignment of the intervention start point. Unless stated otherwise, numerical summaries refer to the flat-baseline, equal-allocation conditions; full results for all 218 conditions are reproducible from the archived output.

Study 1: Level Change

Type I error rates

Table 1 reports Type I error rates under the pure null condition ($\beta_2 = 0$, flat baseline, equal allocation). Under independence ($\rho = 0$), piecewise regression and the Bayesian GLM remained within Bradley's (1978) liberal interval. With increasing autocorrelation both inflated substantially, reaching .27 to .31 at $N \geq 30$ under $\rho = 0.6$. Inflation was not attenuated by longer series but increased with them, because a longer series under positive autocorrelation supplies more nominal but not more effective information. This pattern confirms that without explicit modeling of the error structure, parametric methods detect "effects" that are artifacts of serial dependence. The randomization test maintained Type I error rates below .05 in every condition (.000 at $N = 15$, .017 to .020 at $N = 30$, .006 to .015 at $N = 60$). Read against the attainable levels of .000, .042, and .037, the $N = 15$ rate is exactly what a valid test must produce, whereas the $N = 30$ and $N = 60$ rates fall well below their attainable levels: conservatism beyond discreteness, whose source Study 4 identifies.

Table 1

Type I Error Rates for Level Change Under the Pure Null Condition

N	*ρ*	NAP	Piecewise	Bayesian GLM	Rand. test	Attainable α
15	0.0	.053	.051	.068	.000	.000
15	0.3	.127	.081	.116	.000	.000
15	0.6	.288	.148	.174	.000	.000
30	0.0	.047	.065	.069	.017	.042
30	0.3	.143	.124	.131	.020	.042
30	0.6	.312	.266	.277	.020	.042
60	0.0	.051	.064	.067	.009	.037
60	0.3	.127	.128	.138	.006	.037
60	0.6	.307	.312	.313	.015	.037

Note. N = total series length; ρ = lag-1 autocorrelation; Rand. test = randomization test. All entries are proportions of 1,000 replications with $\beta_2 = 0$, flat baseline, and equal phase allocation. NAP rates are based on Mann–Whitney *p*-values back-computed from the normal approximation to the Wilcoxon null distribution. Attainable α is the largest significance level attainable by the randomization test at or below .05 given the number of admissible start points, and is the correct benchmark for that column only. Bradley's (1978) liberal interval is [.025, .075]. Tau-U admits no *p*-value and Hedges' *g* was not used as a test, so neither appears here.

Statistical power

At the nominal .05 level the Bayesian GLM exhibited the highest sensitivity, marginally above piecewise regression, and the randomization test the lowest; at $N = 15$ its power was exactly 0 for every effect size, the minimum attainable p-value exceeding .05. These nominal comparisons, however, are not comparisons of sensitivity, because the methods were not operating at the same actual significance level. Table 2 therefore reports size-adjusted power, and the reordering is complete. At $N = 30$, $\rho = 0.6$, and $\beta_2 = 1.0$, nominal power was .348 for piecewise regression, .364 for the Bayesian GLM, and .076 for the randomization test; the corresponding size-adjusted values were .106, .107, and .164. At $N = 60$, $\rho = 0.3$, and $\beta_2 = 1.0$, nominal power of .456, .459, and .167 became size-adjusted power of .322, .324, and .413. Once actual size was equated, the randomization test was at least as powerful as both parametric methods in every flat-baseline, equal-allocation level-change condition at $N = 30$ and $N = 60$ (Table 2), and the three methods were indistinguishable in those conditions at $N = 15$ (size-adjusted power of .079 to .156). The full-grid calibration bounds this reversal: with equal allocation the largest departure was .001 (trending baseline, within simulation error); under the imbalanced allocation the advantage narrowed and in eight of twenty-four $N \geq 30$ cells reversed, by up to .09. The apparent power advantage of the parametric methods is therefore not a property of the estimators but an accounting consequence of rejecting more often. Size-adjusted power for the two parametric methods differed by at most .005 in any level-change cell, a first indication that the Bayesian implementation contributed little beyond the OLS fit.

Table 2

Nominal and Size-Adjusted Power for Level Change

N	ρ	β_2	Piecewise		Bayesian GLM		Rand. test	
			Nominal	Adjusted	Nominal	Adjusted	Nominal	Adjusted
30	0.0	0.5	.088	.077	.095	.077	.035	.121
30	0.0	1.0	.275	.250	.293	.252	.140	.365

30	0.3	0.5	.169	.078	.183	.077	.034	.094
30	0.3	1.0	.288	.162	.315	.163	.087	.228
30	0.6	0.5	.266	.054	.275	.054	.027	.088
30	0.6	1.0	.348	.106	.364	.107	.076	.164
60	0.0	0.5	.171	.146	.182	.146	.071	.243
60	0.0	1.0	.472	.423	.478	.425	.257	.581
60	0.3	0.5	.221	.128	.228	.127	.035	.160
60	0.3	1.0	.456	.322	.459	.324	.167	.413
60	0.6	0.5	.337	.061	.342	.061	.023	.077
60	0.6	1.0	.451	.163	.456	.163	.079	.201

Note. Nominal = proportion of replications with $p < .05$. Adjusted = size-adjusted power, that is, the proportion of replications whose *p*-value fell at or below the 5th percentile of the *p*-value distribution obtained under the corresponding null condition for the same method and design cell. Size-adjusted power equates the actual Type I error rate across methods and is reported as a diagnostic; it is not a procedure available to an applied analyst, because the calibration requires knowledge of the true autocorrelation. Flat baseline, equal phase allocation, 1,000 replications per cell. Rows for $N = 15$ are omitted: nominal power for the randomization test is 0 by construction, and size-adjusted power is .079 to .156 for all three methods (see text). Achieved Type I error at the calibrated critical value was .050 to .053 in every cell and method; the randomization test's calibrated critical values, .112 to .148, lie well above .05, a further expression of the systematic design's conservatism.

Coverage probability

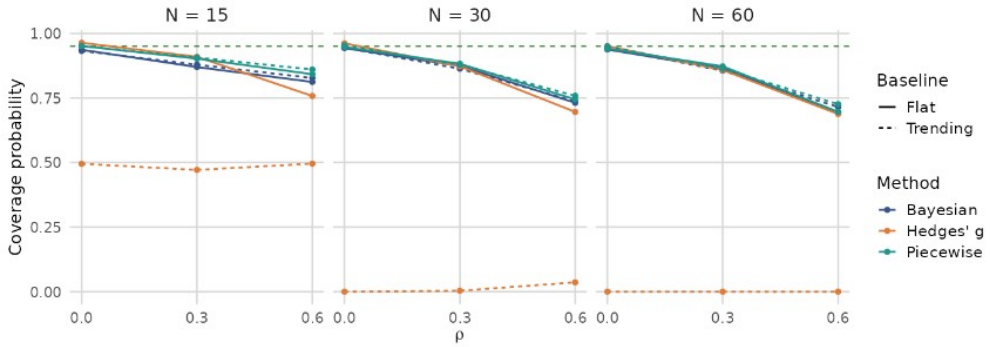
Coverage probability for Hedges' *g* depended jointly on autocorrelation and baseline trend, which produce distinct failure modes. Under a flat baseline, coverage was near nominal at $\rho = 0$ (.951 to .964) and degraded monotonically with serial dependence, to .858 to .909 at $\rho = 0.3$ and .688 to .758 at $\rho = 0.6$: a standard-error problem: intervals correctly centered but too narrow. Under an upward baseline trend ($\beta_1 = 0.2$), coverage collapsed, to .47 to .50 at $N = 15$ and .000 to .036 at $N = 30$ and $N = 60$: a bias problem: the unadjusted standardized mean difference absorbs the trend, which with 15 observations per phase displaces the phase means by 3.0 error standard deviations, so the estimator no longer estimates β_2. The previous version of this article reported coverage for Hedges' *g* as falling below .50 across all conditions; that figure was an average across the two baseline-trend conditions and obscured this distinction. Coverage for the level parameter from piecewise regression and the Bayesian GLM (Figure 1) was unaffected by baseline trend, because both specifications include a trend term, and fell from approximately .95 at $\rho = 0$ to .69 to .86 at $\rho = 0.6$. Coverage was not computed for NAP

or Tau-U (see Evaluation Criteria), and neither appears in Figure 1; NAP's inferential performance is instead evaluated through its Type I error rate.

Under the pure null condition ($\beta_2 = 0$, no baseline trend), NAP maintained nominal Type I error at ρ = 0 (.047 to .053 across series lengths), with inflation to .127 to .143 at $\rho = 0.3$ and .288 to .312 at $\rho = 0.6$, the latter far exceeding Bradley's (1978) liberal upper bound of .075. Under a trending baseline ($\beta_1 = 0.2$, $\beta_2 = 0$), false-positive rates were .589 to .700 at $N = 15$ and .984 to 1.000 at $N = 30$ and $N = 60$. These findings demonstrate that NAP cannot distinguish genuine intervention effects from baseline trend, rendering it unsuitable for inferential use whenever autocorrelation or baseline slope is non-negligible. The trend result is the more serious: it worsens rather than attenuates with sample size, so at typical series lengths a Mann–Whitney test on NAP under even a modest secular trend is all but certain to reject.

Figure 1

Coverage Probability for Level Change by Baseline Trend (Equal Allocation)



Study 2: Slope Change

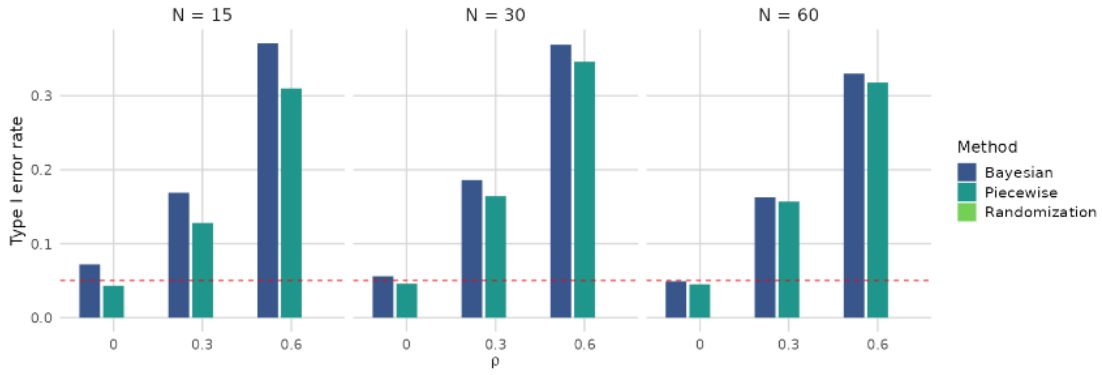
Type I error rates

Piecewise regression and the Bayesian GLM again showed inflated Type I error for the slope parameter (Figure 2), more severe than for the level parameter: .128 to .186 at $\rho = 0.3$ and .310 to .371 at $\rho = 0.6$ across the two methods (reaching .389 under the trending-baseline

null). The randomization test based on the difference in fitted slopes rejected in 1 of the 18,000 null replications across the 18 slope null conditions examined at equal allocation. This is stronger than conservatism: under the null the *p*-values were centered near .70 rather than distributed uniformly, indicating a reference distribution structurally misaligned with the sampling distribution of the observed statistic. Study 4 shows this behavior is produced jointly by the systematic design and the unstandardized slope-difference statistic, and that it is repairable.

Figure 2

Type I Error Rates for Slope Change (Flat Baseline, Equal Allocation)



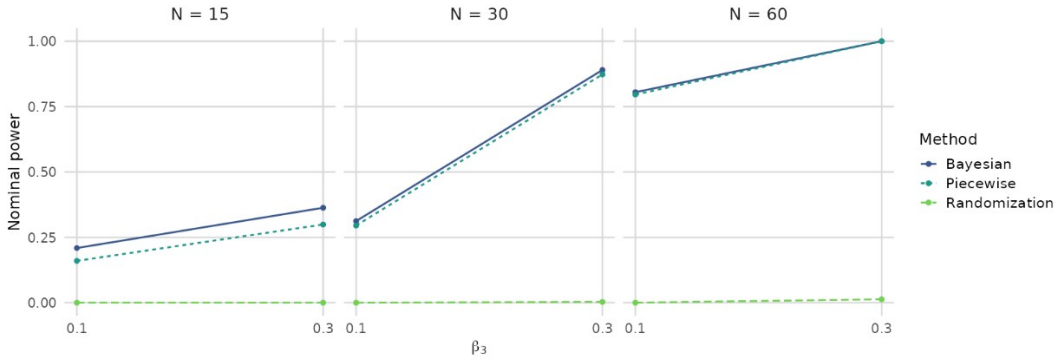
Power to detect trends

Detecting gradual slope changes proved substantially harder than detecting level shifts (Figure 3). For a small slope change ($\beta_3 = 0.1$) at $N = 30$ with equal allocation and $\rho = 0.3$, nominal power was .295 for piecewise regression, .312 for the Bayesian GLM, and .000 for the randomization test. Size-adjusted power narrows but does not close this gap: at $N = 30$ with $\beta_3 = 0.3$, it was .932 for both parametric methods and .458 for the randomization test at $\rho = 0$, and .343 versus .145 at $\rho = 0.6$. Unlike the level-change case, the sensitivity deficit for trend detection is not solely an artifact of size; it reflects the statistic. An unstandardized difference

between fitted slopes has a null variance that depends strongly on where the series is divided, a dependence Study 4 quantifies. Variance-stabilized alternatives exist (Levin et al., 2021), and the present results evaluate the unstandardized statistic rather than randomization inference for trend in general.

Figure 3

Statistical Power for Slope Change (ρ = 0.3, Flat Baseline, Equal Allocation)

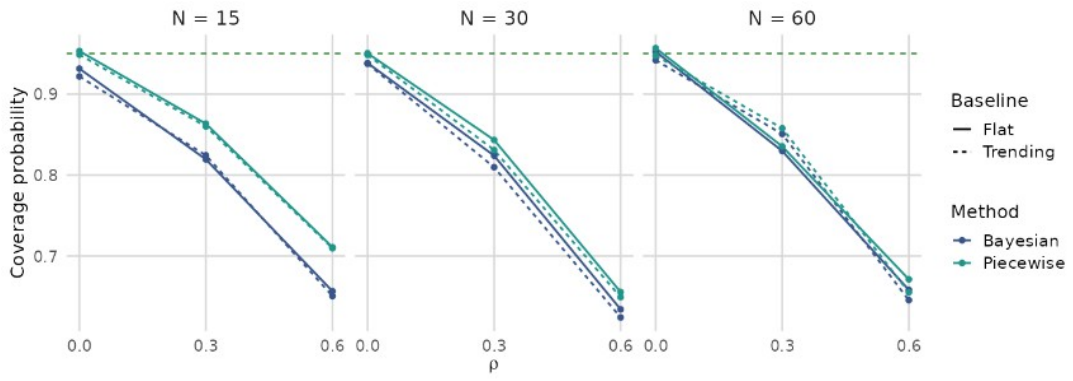


Coverage probability

Consistent with the level-change findings, coverage for slope estimates in piecewise regression and the Bayesian GLM deteriorated with increasing autocorrelation (Figure 4), from .92 to .96 at ρ = 0, to .81 to .86 at ρ = 0.3, and to .62 to .71 at ρ = 0.6. Unlike coverage for Hedges' *g*, slope coverage was essentially unaffected by baseline trend, because the piecewise specification includes a baseline trend term and therefore estimates β_3 rather than a quantity confounded with it. NAP and Tau-U are not evaluated for slope change because they are ordinal indices of phase dominance and contain no parameter corresponding to a change in slope; the question of coverage for such a parameter does not arise within these frameworks.

Figure 4

Coverage Probability for Slope Change by Baseline Trend (Equal Allocation)



Study 3: Combined Effects

Parameter recovery

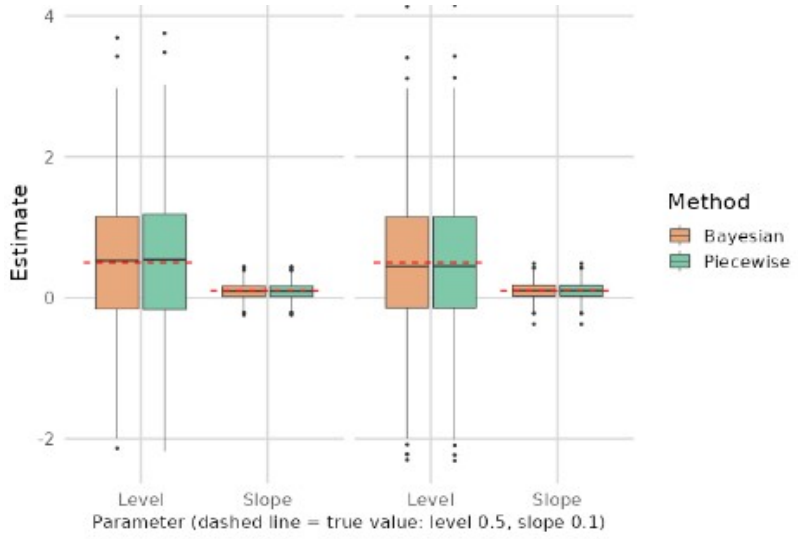
With simultaneous level ($\beta_2 = 0.5$) and slope ($\beta_3 = 0.1$) effects, both piecewise regression and Bayesian GLM demonstrated near-zero bias for each parameter separately (Figure 5). For the level component β_2: bias ≈ +0.03 (Piecewise) and +0.02 (Bayesian) under a flat baseline, and ≈ −0.04 (both) under an upward baseline. For the slope component β_3: bias ≈ −0.008 under a flat baseline and smaller than .001 in absolute value under a trending baseline, for both methods. However, sampling variability was substantial: RMSE for $\beta_2 \approx 0.94$ across methods (compared to a true value of 0.5, ratio ≈ 1.88), and RMSE for $\beta_3 \approx 0.12$ (compared to a true value of 0.1, ratio ≈ 1.16). Coverage probability fell below the nominal 0.95 for both parameters under both baseline conditions; under a flat baseline, for β_2, coverage ≈ .877 (Piecewise) and .864 (Bayesian), and for β_3, coverage ≈ .836 (Piecewise) and .820 (Bayesian) (trending baseline: .815 to .865).

These coverage deficits are consistent with the illusion-of-precision pattern documented in Studies 1 and 2, and confirm that the autocorrelation-induced narrowing of intervals extends to the multiparameter case. Type I error rates and statistical power cannot be evaluated for Study 3 because the simulation includes only a non-null combined-effect condition (true $\beta_2 = 0.5$, β_3

= 0.1); no null condition was included by design, and evaluating a Type I error rate requires one. NAP and Tau-U were not evaluated for combined effects because ordinal indices of phase dominance contain no parameters corresponding to separate level and slope components, so the question of separately identifying them does not arise.

Figure 5

Parameter Recovery in the Combined Effect Model by Baseline Trend



Study 4: Randomized Versus Systematic Assignment of the Intervention Start Point

Studies 1 and 2 documented two anomalies that the systematic design makes impossible to interpret: Type I error below the attainable level for the level-change statistic, and no rejections at all for the slope-change statistic. A supplementary simulation determined how much of each is attributable to the absence of randomization and how much to the test statistic, comparing two conditions under the pure null at each combination of series length (15, 30, 60) and autocorrelation (0, 0.3, 0.6), with 20,000 replications per cell and the reference distribution enumerated exactly. In the systematic condition the realized start point was fixed at the

balanced division, as in the equal-allocation conditions of Studies 1 through 3; in the randomized condition it was drawn uniformly from the admissible set for each replication. Because the data are invariant to the start point under the sharp null hypothesis, this construction reproduces a genuinely randomized start-point design exactly, with all other features of the test held constant.

Table 3 reports the results. For the level-change statistic, randomization restored the test to its attainable size: Type I error in the randomized condition was .041 to .042 at $N = 30$ and .035 to .039 at $N = 60$, against attainable levels of .042 and .037, with no discernible effect of autocorrelation. In the systematic condition the corresponding rates were .015 to .020 and .006 to .011, between one half and one sixth of the attainable level. At $N = 15$ both conditions produced .000, as they must, since no attainable level exists at or below .05.

For the slope-change statistic the contrast is starker: randomized rates were again at the attainable level (.042 to .043 at $N = 30$, .036 to .037 at $N = 60$), whereas the systematic condition produced .000 to .001. The mechanism is visible in the null variance of the statistic across the randomization set. At $N = 30$ and $\rho = 0.3$, the standard deviation of the mean-difference statistic ranges from 0.51 at the balanced division to 0.77 at the most extreme admissible division, a ratio of approximately 1.5; for the unstandardized slope-difference statistic the range is 0.12 to 0.72, a ratio of approximately 6. Under equal allocation, a systematic design always evaluates the observed statistic at the balanced division, which has the smallest null variance in the admissible set, against a reference distribution dominated by divisions of much larger variance; the observed value is therefore almost never extreme, and the p-value is inflated, its null median .71 rather than .50. Studentizing the slope-difference statistic by its null standard deviation at each division (known here by construction; in practice

it would require estimation, cf. Levin et al., 2021) raised the systematic-design Type I error from .000 to .019, .022, and .022 at ρ = 0, 0.3, and 0.6, recovering roughly half of the attainable level.

Table 3

Type I Error Rates Under Systematic Versus Randomized Assignment of the Intervention Start Point

Test statistic	*N*	Attainable α	ρ = 0		ρ = 0.3		ρ = 0.6	
			Sys.	Rand.	Sys.	Rand.	Sys.	Rand.
Mean difference	15	.000	.000	.000	.000	.000	.000	.000
Mean difference	30	.042	.015	.042	.015	.042	.020	.041
Mean difference	60	.037	.006	.039	.007	.037	.011	.035
Slope difference	15	.000	.000	.000	.000	.000	.000	.000
Slope difference	30	.042	.000	.042	.000	.043	.001	.043
Slope difference	60	.037	.000	.036	.000	.037	.000	.036

Note. Sys. = systematic design, intervention start point fixed at the balanced division; Rand. = randomized design, start point drawn uniformly from the admissible set for each replication. Reference distributions were enumerated exactly; 20,000 replications per cell; pure null hypothesis. Attainable α is the largest significance level attainable at or below .05 given the number of admissible start points, and is the correct benchmark for every entry. At $N = 15$ no attainable level exists at or below .05, so a rate of .000 is the only possible outcome under either design.

Three conclusions follow. First, the conservatism in Studies 1 and 2 is not an intrinsic property of randomization tests, nor is it explained by the observed allocation belonging to the reference distribution, which is equally true of a randomized design; it is produced by fixing the allocation at a division whose statistic is not exchangeable with the remainder of the reference set. Second, under a randomized design the test attained the largest significance level its discrete reference distribution admits, within Monte Carlo error, at every autocorrelation level and for both statistics at $N \geq 30$; at $N = 15$ no level at or below .05 is attainable under either design. Attaining the exact attainable level irrespective of the error structure is a property no other method evaluated here possesses. Third, the choice of test statistic matters in proportion to how strongly its null variance varies across the randomization set, which for unstandardized slope differences is severe.

Summary of Findings

The simulation results are summarized as follows.

1. Piecewise regression and the Bayesian GLM without autocorrelation modeling did not control the Type I error rate under serial dependence, reaching .27 to .31 (level) and .31 to .37 (slope) at $\rho = 0.6$, with coverage falling to approximately .69 and .62; point estimates were approximately unbiased. Because their actual significance level is not .05, their nominal power is not a measure of sensitivity; at equated size they were the least sensitive methods for level change.
2. Randomization tests were unaffected by autocorrelation. Under the systematic design of Studies 1 through 3 they were markedly conservative, and at $N = 15$ their power was exactly 0 by construction. Under a randomized start-point design (Study 4) they attained the largest significance level the discrete reference distribution admits at $N \geq 30$, and at equated size their sensitivity to level change equaled or exceeded that of both parametric methods in the flat-baseline, equal-allocation conditions (reversals up to .09 under imbalanced allocation). Their sensitivity to slope change was low for the unstandardized statistic used here.
3. Nonoverlap indices (NAP and Tau-U) yield rank-based summaries of phase dominance whose point values are little affected by autocorrelation. No interval procedure valid under serial dependence is currently available for either index (Pustejovsky et al., 2024), and none was computed here. They contain no parameters corresponding to separate level and slope components, and inferential use of NAP through the Mann–Whitney approximation is invalid under either autocorrelation or baseline trend.

Taken together, these results do not describe a menu of methods among which researchers may trade validity for sensitivity. Of the four methods for which a Type I error rate is defined, three failed to control it under conditions that are common in applied SCED data, and a test that does not control its false-positive rate does not become admissible by being powerful. The practical question is therefore not which trade-off to accept but how to obtain a valid test with acceptable sensitivity, and the results point to the design rather than to the analysis for the answer.

Discussion

Through a systematic evaluation of six statistical methods across 218 simulation conditions, together with a supplementary comparison of systematic and randomized start-point assignment, the present findings clarify where commonly used analytical techniques succeed and where they fail. The overarching conclusion is that no single method performs optimally across all criteria. The results could be summarized as a trade-off between the robustness of randomization-based inference and the sensitivity of model-based inference. That summary does not survive closer examination: equating actual significance levels removes most of the sensitivity advantage of the model-based methods, and introducing randomization into the design removes the conservatism of the randomization test.

The Destabilizing Role of Autocorrelation and the Illusion of Precision

A central contribution of this simulation is the empirical documentation of the severe consequences of unmodeled autocorrelation for parametric inference in SCED data. Consistent with earlier warnings in the methodological literature (e.g., Ferron, 2002; Manolov & Solanas, 2008; Solomon, 2014), the present results confirm that piecewise regression and Bayesian generalized linear models, when implemented without explicit autocorrelation modeling,

produce Type I error rates of .27 to .31 for the level parameter and .32 to .37 for the slope parameter at $N \geq 30$ under high autocorrelation ($\rho = 0.6$).

The analysis of coverage probability extends this picture. Coverage has been examined for individual SCED methods in prior work (Baek et al., 2020; Chen et al., 2025; Li et al., 2025; Natesan Batley & Hedges, 2021), and the present study follows that line rather than opening it. What the present design adds is the evaluation of coverage alongside Type I error and power under identical data-generating conditions for methods from three paradigms, with an explicit account of which methods admit which criterion. The present results demonstrate that at ρ = 0.6, coverage for piecewise regression and the Bayesian GLM dropped to approximately .69 for the level parameter and .62 for the slope parameter, substantially below the nominal .95. The degradation reflects systematic overconfidence: under positive autocorrelation the effective sample size is smaller than the nominal one, compressing standard errors. The Hedges' *g* results show a second mechanism producing the same symptom: under a trending baseline coverage collapsed to essentially zero at $N \geq 30$, reflecting not an underestimated standard error but a misspecified estimand, an unadjusted mean difference absorbing the trend into its numerator so that the interval is correctly wide but wrongly centered. The distinction matters because widening intervals repairs the first failure and cannot repair the second.

Since Huitema's (1985) claim that autocorrelation is negligible, the debate has continued (Tanious & Onghena, 2026). Large-scale reviews of published SCED data document non-trivial autocorrelation in applied settings (Pustejovsky et al., 2023; Shadish & Sullivan, 2011; Solomon, 2014), indicating that independence is frequently violated in practice.

However, it must be acknowledged that accurately estimating autocorrelation from short time series is itself a difficult statistical problem. Lag-1 autocorrelation estimates from series of

10 to 20 observations are highly variable and often biased (Huitema & McKean, 1998; Shadish et al., 2013). This estimation difficulty creates a practical dilemma: the very conditions under which autocorrelation is most damaging (short series) are also the conditions under which it is hardest to detect and quantify. The present findings therefore support a precautionary approach.

Type I Error Control as a Prerequisite

The second major finding concerns what has conventionally been described as a trade-off between Type I error control and statistical power. The randomization test was the only method that maintained an admissible Type I error rate across all conditions, consistent with theory (Edgington, 1980; Edgington & Onghena, 2007) and prior simulation evidence (Bouwmeester & Jongerling, 2020; Ferron et al., 2003; Ferron & Sentovich, 2002; Michiels & Onghena, 2019). Its insensitivity to autocorrelation follows from the reference distribution's construction: the data remain in temporal order and only the phase-change location varies, so the serial structure is identical under every division. Whether the resulting *p*-value can be interpreted as a *p*-value, however, depends on something the computation cannot supply. A randomization test is justified by the randomization actually performed in the design; the enumeration of data divisions has inferential content only when one of those assignments was in fact selected at random (Edgington, 1980; Ferron et al., 2003). Where the start point was fixed by the investigator, no assignment mechanism exists, and the term design-based inference, used in the previous version of this article, was not warranted. That term has been removed throughout.

The consequences of that absence are quantified in Study 4, and they are not what the previous version of this article proposed. It was suggested there that a systematic design produces conservative Type I error rates because the observed allocation is by definition always included in the reference distribution. That explanation is incorrect: the realized

assignment belongs to the reference set in a randomized design as well, since the reference set is precisely the set of assignments the design could have produced. The conservatism instead arises because, under equal allocation, a systematic design always evaluates the observed statistic at the same balanced division, whose null variance is the smallest in the admissible set, against a reference distribution pooling divisions of much larger variance; the observed value is therefore rarely extreme. When the realized start point is drawn at random, so that the realized division is a uniform draw from the admissible set, the test attains the largest significance level its discrete reference distribution admits at every autocorrelation level examined (Table 3). The conservatism documented in Studies 1 through 3 is thus a property of the joint use of a randomization test and a non-randomized design, eliminated not by a different analysis but by a different design.

The cost of that conservatism was substantial, and at $N = 15$ absolute: power was exactly 0, because the minimum attainable p-value of $1/9 = .111$ exceeds $\alpha = .05$ by construction, a consequence of the restricted randomization space noted in prior work (Levin et al., 2018; Michiels & Onghena, 2019). At $N = 30$ and $N = 60$, however, the power deficit at the nominal level proved largely an artifact of the comparison. Once actual size was equated, the randomization test was at least as sensitive to level change as either parametric method in every flat-baseline, equal-allocation condition at those series lengths, and its advantage persisted at every autocorrelation level examined. The parametric methods were not detecting more real effects; they were rejecting more often.

The one genuine sensitivity deficit that remains is trend detection: size-adjusted power for the randomization test built on an unstandardized slope difference was roughly half that of the parametric methods. Study 4 traces this to the strong dependence of that statistic’s null

variance on the phase division and shows that studentizing recovers roughly half of the attainable level. The appropriate conclusion concerns test-statistic selection rather than randomization inference for trend: unstandardized slope differences are a poor choice, and variance-stabilized alternatives are available (Levin et al., 2017, 2021). This bears directly on SCEDs in which effects are expected to be gradual, such as skill-acquisition programs or interventions for chronic conditions.

It is worth stating plainly what these results do and do not license. They do not license a choice between a valid test and a powerful one. Controlling the false-positive rate is a precondition for a significance test in the same sense that internal validity is a precondition for external validity: a procedure that identifies an inert intervention as effective in more than a quarter of samples is not sensitive but uninformative, and the frequency with which it also flags effective interventions is not a compensating virtue. What they do license is a redirection of the question: given that a valid test is required, how is adequate sensitivity to be obtained from one? The answer indicated by Study 4 lies in the design rather than in the analysis.

The Role of the Nonoverlap Indices

The nonoverlap indices evaluated here, NAP and Tau-U, occupy a distinct niche. The present results demonstrate that inferential use of NAP through the Mann–Whitney normal approximation is untenable under SCED conditions: Type I error reached .14 at $\rho = .3$ and .31 at $\rho = .6$, far outside Bradley's (1978) liberal robustness interval of [.025, .075] at $\alpha = .05$.

Under baseline trend, false-positive rates approached 1.0, a consequence not of the normal approximation but of NAP's ordinal structure, which cannot distinguish intervention-attributable phase separation from trend-induced separation. For Tau-U with baseline correction, no valid sampling distribution exists (Pustejovsky et al., 2024), precluding interval-

based inference entirely. These indices are therefore best understood as descriptive summaries of ordinal dominance between phases, quantifying how consistently intervention-phase observations exceed baseline-phase observations, rather than as measures of the magnitude of a difference or as tools for formal hypothesis testing under serial dependence (Pustejovsky, 2019).

Point estimates from NAP and Tau-U were largely unaffected by autocorrelation. No interval estimate was computed for either index; the practical implication is that they can be reported descriptively but cannot at present be accompanied by an interval trustworthy under serial dependence.

Two additional considerations are relevant. First, as Carter (2013) and Manolov and Tanious (2025) have noted, nonoverlap indices are subject to a ceiling effect that limits their sensitivity in distinguishing among large effects (see also Manolov et al., 2011; Vannest et al., 2018): because they measure dominance rather than distance, they saturate once separation is complete, constraining fine-grained comparison in meta-analytic contexts. Second, the interpretive challenges of Tau-U are documented (Brossart et al., 2018), and the denominator ambiguity noted in the Method section compounds them; baseline-corrected Tau (Tarlow, 2017) may offer improved properties. Despite these limitations, nonoverlap indices remain valuable as accessible, distribution-free summaries of treatment impact.

Toward an Integrated Analytic Strategy

The findings do not support choosing among methods according to a preferred balance of validity and sensitivity, because three of the four methods yielding a *p*-value do not supply valid inference under common SCED conditions. They support a division of labor: hypothesis testing, quantification of change, and estimation of specific model parameters are distinct tasks

with distinct requirements, and the strategy below assigns each to the method that meets them rather than treating the methods as interchangeable competitors.

Confirmatory hypothesis testing

For the question of whether an intervention produced a change distinguishable from chance, the requirement is a test that controls its false-positive rate; among the methods evaluated, only the randomization test did so under serial dependence. Where the start point has been randomized, the test provides that control exactly, without any assumption about the error structure, and the size-adjusted comparisons (Table 2) suggest no sensitivity penalty for this validity, although randomization-test power under a randomized design was not directly evaluated (see Limitations). Where the start point was not randomized, the position should be stated candidly: the enumeration of data divisions can still be computed, but it lacks a design-based justification and is markedly conservative in a way that depends on the test statistic and the phase division used; such an analysis is best reported as an exploratory sensitivity check rather than as a confirmatory test. When a randomization test in a short series yields a non-significant result, researchers should resist substituting a less robust method in pursuit of significance; the result should be interpreted in light of known power limitations and supplemented with descriptive information.

Effect quantification

To describe the magnitude of observed change, researchers should report one or more nonoverlap indices alongside a model-based estimate: NAP as a probability-based summary of phase separation, Tau-U for additional control of monotonic baseline trend. Neither should carry an interval computed under an independence assumption (Pustejovsky et al., 2024); reporting the point estimate with an explicit statement to that effect is preferable to an interval

that cannot be interpreted.

Parameter estimation for complex effects

When the research question involves disentangling simultaneous level and slope changes or estimating specific model parameters, regression-based approaches remain essential. The present findings argue against ordinary least squares estimation without autocorrelation modeling. Researchers should employ generalized least squares approaches (Maggin et al., 2011; Swaminathan et al., 2014b) or multilevel models accommodating autocorrelation across participants (Baek & Ferron, 2013). The design-comparable standardized mean difference (Hedges et al., 2012, 2013) represents another principled option when cross-design synthesis is a goal. In the present simulation the Bayesian GLM with weakly informative priors was practically indistinguishable from OLS piecewise regression: size-adjusted power differed by at most .005 (level change) and .008 (slope change) in any design cell, bias by at most .05, and coverage by at most .06 (largest differences at $N = 15$). Where the two differed in Type I error, the Bayesian implementation was always the more liberal, exceeding piecewise regression in every one of the 72 null conditions, by up to .04 for the level parameter and .07 for the slope parameter. This is the expected result: priors carrying no substantive information about effect size cannot supply information the data do not contain, and the principal argument for Bayesian analysis in single-case research is precisely the ability to incorporate such information, for example from meta-analytic evidence (Natesan & Hedges, 2017; Shadish et al., 2013). A weakly informative Bayesian analysis of an AB design therefore adds computational and interpretive complexity, and an additional opportunity for misinterpretation, in exchange for no inferential gain, and a simpler analysis that applied researchers are more likely to interpret correctly is preferable. Bayesian methods are

recommended only where genuinely informative priors can be defended, or where the estimation problem requires them, as in unknown change-point models (Natesan & Hedges, 2017).

Practical Implications for Research Design

Beyond analytic choices, the present findings carry design implications, and these are the more consequential. First, researchers should plan for adequate series length: at $N = 15$, randomization-test power is zero by construction and all methods struggle to detect slope changes, consistent with design standards emphasizing sufficient data density (What Works Clearinghouse, 2022).

$N = 30$, with 15 observations per phase, is a meaningful practical threshold: 24 admissible start points and an attainable level of .042, so significance is possible, whereas at $N = 15$ no result can be significant at $\alpha = .05$. Investigators for whom formal inference is a priority should treat $N = 30$ as a practical minimum, noting that the requirement is driven by the number of admissible assignments rather than sample size as such: randomizing the start point within a wider admissible window achieves a smaller attainable level at the same series length.

The second and more important recommendation concerns randomization of the intervention start point. Study 4 shows that this single design feature converts the randomization test from a markedly conservative procedure into one that attains exactly the largest significance level its discrete reference distribution admits, at every autocorrelation level examined and for both test statistics at $N \geq 30$, without any distributional assumption and without estimating the autocorrelation. No analytic remedy evaluated here achieves the same result, and randomization additionally grounds the inference in the assignment mechanism

actually used (Edgington, 1980; Edgington & Onghena, 2007; Tanious & Onghena, 2021). Where clinical or educational constraints preclude randomization, investigators should say so; if the enumeration is computed at all, it should be labeled an exploratory sensitivity check rather than presented as a test, since without randomization the *p*-value has no design-based warrant and is, on the present evidence, likely to be substantially conservative. Designs that randomize within a restricted window of clinically acceptable start points retain the inferential warrant while limiting the delay to intervention. A further refinement bears on such recommendations. The Type I error rates reported in Study 4 are marginal rates, averaged over the admissible assignment set; the corresponding rates conditional on a particular realized division need not be equal, because not all admissible start points are associated with the same Type I error rate. Evaluating the test separately for each admissible randomization, an approach termed conditional (Sierra et al., 2005) or data-division-specific (Manolov et al., 2010; see also Manolov & Solanas, 2009, on the AB design considered here), makes that heterogeneity visible; the systematic condition of Study 4 is in effect one such conditional evaluation, taken at the balanced division, and its marked conservatism shows how far a single division can depart from the marginal rate. A data-division-specific characterization of restricted randomization windows is a natural extension of the present results.

Limitations and Future Directions

Several limitations of the present study should be noted. First, data were generated using an AR(1) process with normally distributed errors. Behavioral time series may exhibit more complex dependence structures, including MA or ARMA processes and non-stationarity. Future work should examine whether findings generalize to these alternative error structures. The simulation was also limited to continuous outcomes; applied SCED research increasingly

involves count data, proportions, and other non-normal outcomes (Cho et al., 2022; Declercq et al., 2019; Li, 2024; Swan et al., 2020), for which method performance may differ.

Second, the Bayesian implementation employed weakly informative priors and penalized maximum likelihood with a normal approximation to the posterior rather than Markov chain Monte Carlo sampling. This is the implementation most accessible to applied researchers, not the most capable one, and the finding that it performed no better than OLS is a statement about weakly informative priors and approximate posterior summaries, not about Bayesian single-case analysis in general. Fully sampled models with informative priors from meta-analytic evidence (Shadish et al., 2013), models placing a prior on the autocorrelation itself, and unknown change-point models (Natesan & Hedges, 2017) may behave very differently, and sensitivity to prior specification remains an important area for future work.

Third, the simulation focused on the AB phase design, which is the simplest SCED structure and does not, on its own, meet the What Works Clearinghouse (2022) standards for demonstrating experimental control. The limits on generalizing differ by method, as set out in the Method section: for the model-based methods the documented properties attach to a single phase contrast and are expected to carry over qualitatively to replicated designs, subject to the structure a multilevel specification imposes (Baek et al., 2020; Baek & Ferron, 2013; Michiels et al., 2020; Moeyaert et al., 2017); for the randomization test no transfer holds, because each design supports a different randomization scheme of different cardinality, with different attainable significance levels and power (Levin et al., 2018; Onghena, 1992; Onghena & Edgington, 1994, 2005). Direct evaluation within ABAB, multiple-baseline, and alternating-treatments schemes is needed and is not supplied here.

Fourth, a single test statistic was used in each scenario: the absolute difference in phase

means for level change and the absolute difference in fitted slopes for slope change, each two-tailed. A directional statistic would not merely be defensible but the logical choice for applied use, because whether the intervention is intended to increase or to reduce the target behavior is known before the data are collected and only one direction is therapeutically desirable; it would yield uniformly higher power, and Study 4 shows that studentizing the slope statistic, where its division-specific null variability can be estimated, substantially improves the trend test. Test-statistic selection is not a secondary implementation detail here but a determinant of the conclusions, and systematic evaluation of alternatives would be informative (Levin et al., 2017, 2021).

Fifth, no interval estimate was computed for NAP or Tau-U, so the present study contributes no evidence about interval coverage for the nonoverlap indices. It establishes only that inferential use of NAP through the Mann–Whitney normal approximation is invalid under serial dependence or baseline trend. Interval procedures that respect the temporal structure, such as block bootstrap or randomization-test inversion (Michiels et al., 2017), could in principle supply valid intervals and merit development.

Sixth, effects were parameterized in innovation-standard-deviation units, so a fixed effect magnitude corresponds to a smaller standardized mean difference as autocorrelation increases; comparisons among methods within an autocorrelation level are unaffected, but the power decline across levels partly reflects a smaller standardized signal. Seventh, Study 4 was restricted to the null hypothesis and the AB start-point scheme; the power of the randomization test under a randomized design was not evaluated and warrants direct study.

Finally, this study focused on the statistical properties of individual analytical methods in isolation. Formal frameworks for evidence synthesis across multiple analytical methods

within a single SCED represent an important direction for future development.

Conclusion

The transition from visual to statistical analysis in SCED research is a necessary evolution accompanied by substantial analytic challenges. The choice of statistical method is not a minor technical decision; it can fundamentally alter the substantive conclusions drawn from a study. There is no analytic remedy for the combination of a short series and temporally dependent errors, where short refers here to the number of measurement occasions within a single AB comparison rather than to a number of participants, and where the series lengths examined in this study are at or above the medians reported in surveys of published work. Piecewise regression and Bayesian generalized linear models fitted without autocorrelation modeling produced Type I error rates up to .37 and coverage as low as .62; their apparent sensitivity advantage did not survive equating actual significance levels. Nonoverlap indices remain useful descriptive summaries of ordinal dominance but are not valid instruments of inference under serial dependence or baseline trend. Randomization tests alone controlled the false-positive rate; under a randomized start-point design they attained the largest significance level their discrete reference distribution admits, with no distributional assumption.

The resolution lies less in the selection of an analysis than in the design of the study. Randomizing the intervention start point, wherever circumstances permit, secures a test that attains its intended significance level without estimating the autocorrelation, with no indication from the size-adjusted comparisons of a sensitivity cost relative to the parametric alternatives at the same false-positive rate, though randomized-design power awaits direct evaluation. Where randomization is genuinely impossible, that should be reported and inferential claims qualified accordingly. Alongside the primary test, nonoverlap indices should be reported to characterize

the magnitude of change, and autocorrelation-aware regression used for estimating specific level and slope parameters. Throughout, researchers should be explicit about their methods' limitations, particularly the likely underestimation of standard errors under serial dependence.

References

Baek, E., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2020). Brief research report: Bayesian versus REML estimations with noninformative priors in multilevel single-case data. *The Journal of Experimental Education, 88*(4), 698–710. https://doi.org/10.1080/00220973.2018.1527280

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods, 45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

Baek, E., & Ferron, J. M. (2021). Bayesian analysis for multiple-baseline studies where the variance differs across cases in OpenBUGS. *Developmental Neurorehabilitation, 24*(2), 130–143. https://doi.org/10.1080/17518423.2020.1858455

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's tau and Tau-U for single-case experimental designs. *Cogent Psychology, 5*(1), Article 1518687. https://doi.org/10.1080/23311908.2018.1518687

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data: What they tell us and what they don't. *Behavior Modification, 37*(3), 378–390. https://doi.org/10.1177/0145445513476609

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*(4), 387–400. https://doi.org/10.1177/002246698501900404

Chen, Y. K., Yang, T. R., Chen, L. T., Hsieh, C. Y., Cheng, C., Wu, P. J., & Peng, C. Y. J. (2025). Improving applications of a design-comparable effect size in single-case designs. *Behavior Research Methods, 57*(10), Article 279. https://doi.org/10.3758/s13428-025-02715-1

Cho, S. J., Naveiras, M., & Barton, E. (2022). Modeling multivariate count time series data with a vector Poisson log-normal additive model: Applications to testing treatment effects in single-case designs. *Multivariate Behavioral Research, 57*(2–3), 422–440. https://doi.org/10.1080/00273171.2020.1860732

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*(6), 828–839. https://doi.org/10.1177/0145445508318606

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods, 51*(6), 2477–2497. https://doi.org/10.3758/s13428-018-1091-y

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*(4), 573–579. https://doi.org/10.1901/jaba.1979.12-573

de Vries, R. M., Hartogs, B. M., & Morey, R. D. (2015). A tutorial on computing Bayes factors for single-subject designs. *Behavior Therapy, 46*(6), 809–823. https://doi.org/10.1016/j.beth.2014.09.013

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18*(2), 165–185. https://doi.org/10.1037/a0031037

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*(3), 235–251. https://doi.org/10.3102/10769986005003235

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011814

Ferron, J. M. (2002). Test of the null hypothesis of no autocorrelation in single-case phase designs. *The Journal of Experimental Education, 70*(4), 297–308. https://doi.org/10.1080/00220970209599511

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

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*(4), 493–510. https://doi.org/10.1037/a0037038

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

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods, 3*(3), 224–239. https://doi.org/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*(4), 324–341. https://doi.org/10.1002/jrsm.1086

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*(2), 107–118.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*(1), 104–116. https://doi.org/10.1037/1082-989X.3.1.104

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38–58. https://doi.org/10.1177/00131640021970358

Joo, S. H., & Ferron, J. M. (2019). Application of the within-and between-series estimators to non-normal multiple-baseline data: Maximum likelihood and Bayesian approaches.

Multivariate Behavioral Research, *54*(5), 666–689. https://doi.org/10.1080/00273171.2018.1564877

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. *The Journal of Experimental Education*, *87*(1), 55–74. https://doi.org/10.1080/00220973.2017.1409181

Kazdin, A. E. (2021). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. What Works Clearinghouse. https://files.eric.ed.gov/fulltext/ED510743.pdf

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

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

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2021). Investigation of single-case multiple-baseline randomization tests of trend and variability. *Educational Psychology Review*, *33*(2), 713–737. https://doi.org/10.1007/s10648-020-09549-7

Li, H. (2024). Model selection of GLMMs in the analysis of count data in single-case studies: A Monte Carlo simulation. *Behavior Research Methods*, *56*(7), 7963–7984. https://doi.org/10.3758/s13428-024-02464-7

Li, H., Luo, W., Baek, E., Thompson, C. G., & Lam, K. H. (2025). Multilevel modeling in single-case studies with count and proportion data: A demonstration and evaluation. *Psychological Methods, 30*(4), 815–842. https://doi.org/10.1037/met0000607

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, 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*(3), 301–321. https://doi.org/10.1016/j.jsp.2011.03.004

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*(6), 860–875. https://doi.org/10.1177/0145445508318866

Manolov, R., & Solanas, A. (2009). Problems of the randomization test for AB designs. *Psicológica, 30*(1), 137–154. https://www.redalyc.org/pdf/169/16911973007.pdf

Manolov, R., Solanas, A., Bulté, I., & Onghena, P. (2010). Data-division-specific robustness and power for ABAB designs. *The Journal of Experimental Education, 78*(2), 191–214. https://doi.org/10.1080/00220970903292827

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*(3), 533–545. https://doi.org/10.1016/j.beth.2010.12.003

Manolov, R., & Tanious, R. (2025). Assessing nonoverlap in single-case data: Strengths, challenges, and recommendations. *Journal of Behavioral Education, 34*(4), 869–901. https://doi.org/10.1007/s10864-024-09552-w

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49*(1), 363–381. https://doi.org/10.3758/s13428-016-0714-4

Michiels, B., & Onghena, P. (2019). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*(6), 2454–2476. https://doi.org/10.3758/s13428-018-1084-x

Michiels, B., Tanious, R., De, T. K., & Onghena, P. (2020). A randomization test wrapper for synthesizing single-case experiments using multilevel models: A Monte Carlo simulation study. *Behavior Research Methods, 52*(2), 654–666. https://doi.org/10.3758/s13428-019-01266-6

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760–778. https://doi.org/10.1037/met0000136

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behavior Modification, 38*(5), 665–704. https://doi.org/10.1177/0145445514535243

Natesan, P., & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods, 22*(4), 743–759. https://doi.org/10.1037/met0000134

Natesan Batley, P., & Hedges, L. V. (2021). Accurate models vs. accurate estimates: A simulation study of Bayesian single-case experimental designs. *Behavior Research Methods, 53*(4), 1782–1798. https://doi.org/10.3758/s13428-020-01522-0

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14*(2), 153–171.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21*(1), 56–68. https://doi.org/10.1097/00002508-200501000-00007

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Lawrence Erlbaum Associates.

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods, 24*(2), 217–235. https://doi.org/10.1037/met0000179

Pustejovsky, J. E., Swan, D. M., & Chen, M. (2024, September 6). *Effect size definitions and mathematical details*. SingleCaseES. https://jepusto.github.io/SingleCaseES/articles/Effect-size-definitions.html

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2023). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification, 47*(6), 1423–1454. https://doi.org/10.1177/0145445519864264

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rindskopf, D. M. (2014). Bayesian analysis of data from single-case designs. *Journal of School Psychology, 52*(2), 179–189. https://doi.org/10.1016/j.jsp.2013.11.006

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*(2), 24–33. https://doi.org/10.1177/074193258700800206

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods, 45*(3), 813–821. https://doi.org/10.3758/s13428-012-0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive models to analyze single-case designs. *Journal of School Psychology, 52*(2), 149–178. https://doi.org/10.1016/j.jsp.2013.11.004

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*(2), 140–160. https://doi.org/10.3200/JEXE.73.2.140-160

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*(4), 477–496. https://doi.org/10.1177/0145445513510931

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014a). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*(2), 213–230. https://doi.org/10.1016/j.jsp.2013.12.002

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G., & Smolkowski, K. (2014b). Regression models for the analysis of single case designs. *Neuropsychological Rehabilitation, 24*(3–4), 554–571. https://doi.org/10.1080/09602011.2014.887586

Swan, D. M., Pustejovsky, J. E., & Beretvas, S. N. (2020). The impact of response-guided designs on count outcomes in single-case experimental design baselines. *Evidence-Based Communication Assessment and Intervention, 14*(1–2), 82–107. https://doi.org/10.1080/17489539.2020.1739048

Tanious, R., & Onghena, P. (2021). A systematic review of applied single-case research published between 2016 and 2018: Study designs, randomization, data aspects, and data analysis. *Behavior Research Methods, 53*(4), 1371–1384. https://doi.org/10.3758/s13428-020-01502-4

Tanious, R., & Onghena, P. (2026). On the need to address serial dependency when dealing with data from single-case experimental designs. *Journal of Behavioral Education*. Advance online publication. https://doi.org/10.1007/s10864-026-09624-z

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification, 41*(3), 427–467. https://doi.org/10.1177/0145445516676750

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and N-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*(5), 619–638. https://doi.org/10.1080/09602011.2013.824383

Vannest, K. J., Peltier, C., & Haas, A. (2018). Results of the Tau-U queries from 2014 to 2017. *Journal of Special Education Technology, 33*(4), 269–277. https://doi.org/10.1177/0162643418786025

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook* (Version 5.0). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Yu, H.-T. (2026). *Evaluating statistical methods for single-case designs under serial dependence: Precision, validity and sensitivity* [Data set and code]. Open Science Framework. https://doi.org/10.17605/OSF.IO/SEZ2T