Evaluation & the Health Professions
2022, Vol. 45(1) 36–53

DOI: 10.1177/01632787211071136
journals.sagepub.com/home/ehp
SAGE

Methods for Modeling Autocorrelation and Handling Missing Data in Mediation Analysis in Single Case Experimental Designs (SCEDs)

Emma Somer[1], **Christian Gische**[2], **and Milica Miočević**[1]

Abstract

Single-Case Experimental Designs (SCEDs) are increasingly recognized as a valuable alternative to group designs. Mediation analysis is useful in SCEDs contexts because it informs researchers about the underlying mechanism through which an intervention influences the outcome. However, methods for conducting mediation analysis in SCEDs have only recently been proposed. Furthermore, repeated measures of a target behavior present the challenges of autocorrelation and missing data. This paper aims to extend methods for estimating indirect effects in piecewise regression analysis in SCEDs by (1) evaluating three methods for modeling autocorrelation, namely, Newey-West (NW) estimation, feasible generalized least squares (FGLS) estimation, and explicit modeling of an autoregressive structure of order one (AR(1)) in the error terms and (2) evaluating multiple imputation in the presence of data that are missing completely at random. FGLS and AR(1) outperformed NW and OLS estimation in terms of efficiency, Type I error rates, and coverage, while OLS was superior to the methods in terms of power for larger samples. The performance of all methods is consistent across 0% and 20% missing data conditions. 50% missing data led to unsatisfactory power and biased estimates. In light of these findings, we provide recommendations for applied researchers.

Keywords

mediation analysis, single-case experiment designs, autocorrelation, missing data, small sample sizes

Single-Case Experiment Designs (SCEDs) are a valuable alternative to Randomized Controlled Trials (RCTs) that enable researchers to evaluate the effectiveness of an intervention at the individual level (Kazdin, 2011; Kratochwill et al., 2013; Shadish & Sullivan, 2011). The main goal of SCEDs is to determine whether there is a causal relationship between a treatment and change in the outcome variable of interest (Krasny-Pacini & Evans, 2018; Smith, 2012). For this aim, a small number of participants are repeatedly measured on variables of interest during baseline and intervention phases. Since participants serve as their own control, researchers can obtain detailed information related to changes over time, and intervention effects at the individual level can be estimated (Barlow et al., 2009).

SCEDs are used across various research fields, including occupational therapy, special education, and rehabilitation (Lane et al., 2017; Ritter et al., 2018; Smith, 2012). Given the heterogeneous nature of behavioral and psychological phenomena, SCEDs provide a valuable alternative to group level studies in populations with low incidence rates or in which analyses at a group level may overlook intervention effects present in certain subgroups (Gaynor & Harris, 2008; Maric et al., 2012). Further, the methodology is useful for evaluating a novel intervention prior to a costly and demanding RCT (Jarrett & Ollendick, 2012; Norell-Clarke et al., 2011). Finally, SCEDs present the opportunity for collaboration between clinicians and researchers, unifying research questions that emerge from clinical practice on one hand and research methodology to evaluate these questions on an individual level on the other hand (Geuke et al., 2019).

Examples of SCEDs include the AB design in which a baseline period A is followed by an intervention period B. In $A_1B_1A_2B_2$ designs, also known as a reversal design, the baseline phase (A_1) is followed by the intervention phase (B_1), the withdrawal of treatment (A_2), and the re-introduction of the intervention (B_2). This type of SCED is useful when changes in behavior caused by an intervention are expected to return to baseline levels once treatment is discontinued. Another common design includes the multiple-baseline design in which participants are randomized to different lengths of

[1] Department of Psychology, McGill University, Montreal, QC, Canada
[2] Department of Psychology, Humboldt-Universitätzu Berlin, Berlin, Germany

Corresponding Author:
Emma Somer, McGill University, 2001 Avenue McGill College, Montréal, QC H3A 1G1, Canada.
Email: emma.somer@mail.mcgill.ca

baseline phase A prior to introducing the intervention phase B, taking into account the effects of maturity and passage of time. For the interested reader, an extensive overview of SCEDs is provided by Tate et al. (2016) and Smith (2012).

Given the growing popularity of SCEDs as a rigorous scientific research approach, there have been efforts to establish empirical methods for evaluating the effectiveness of an intervention (Manolov & Moeyaert, 2017). Despite the prevalence of visual analysis, as described by Kratochwill et al. (2010), statistical analysis of SCED data is preferred since it is less prone to bias and subjectivity (Beeson & Robey, 2006). Efforts to empirically validate indices and effect sizes are of particular interest since it is useful to quantify the size of the intervention effect. Nevertheless, the index of choice depends on the aims of the study, and some indices may be better suited than others (Manolov & Moeyaert, 2017). Non-parametric non-overlap indices, such as non-overlap of all pairs (NAP; Parker & Vannest, 2009), improvement rate difference (IRD; Parker et al., 2009), Tau-U (Parker et al., 2011), and the percentage of non-overlapping corrected data (PNCD, Manolov & Solanas, 2009), are useful for measuring the degree of non-overlap between the baseline and treatment data. Descriptive indices, such as the percentage change index (PCI; Hershberger et al., 1999; or percentage reduction data, as referred to by Wendt, 2009), slope and level change (SLC; Solanas et al., 2010), and mean phase difference (MPD; Manolov & Solanas, 2013), quantify the change in level and slope. Parametric approaches are useful for quantifying the treatment effect size and estimating the standard error. Examples of parametric approaches are standardized mean differences (e.g., Cohen's d, Hedge's g; Shadish et al., 2014), regression-based effect sizes (Center et al., 1985; Swaminathan et al., 2014), multilevel modeling (Ferron et al., 2010; Moeyaert et al., 2014), and between-case standardized difference (Hedges et al., 2012, 2013).

The approach examined in this paper relies on regression-based methods, first proposed by Center et al. (1985). Piecewise regression procedures involve fitting separate models for each phase, baseline and intervention, using ordinary least squares regression (OLS). Given that the assumptions of OLS regression hold, such as the assumption that the outcome is continuous, the residuals are homoscedastic and uncorrelated, and the residuals are normally distributed with means of zero in the population, we can obtain unbiased estimates of at least two regression-based effect sizes: an immediate intervention effect (i.e., immediate change in level) and the intervention effect on the time trend. Using an AB design to illustrate the technique (Equation (1)), the intercept, b_0 of the piecewise regression model provides an estimate of the level of the first measurement of the outcome variable in phase A. The regression coefficients provide an estimate for the trend in phase A (i.e., the regression coefficient for the linear time variable, b_1), the change in level at the onset of phase B (i.e., the difference between the intercept of phase B and the predicted score if this were a score in phase A, b_2), and the change in the trend between the phases (i.e., the difference in linear trends between the A and B phases, b_3). The following equation describes the observed score (Y_t) at time t

$$\begin{aligned} Y_t &= b_0 + b_1\, time_t + b_2\, phase_t \\ &\quad + b_3\, phase\_time_t + e_t \end{aligned} \tag{1}$$

Despite the advantages of SCEDs mentioned above, there are several methodological challenges that may prevent the proliferation of SCED methodology in clinical intervention research. A common characteristic of SCEDs is serial dependency among error terms, commonly referred to as autocorrelation. Autocorrelation among consecutive error terms can be modeled via autoregressive (AR) processes, for example, an AR process of order one (AR(1). Autocorrelation is quantified by the parameter rho (ρ), which ranges from -1 to 1 (see Equations (7) and (8) in the section on Autocorrelation Modeling Techniques). Violating the assumption of independence of errors as required by most parametric and non-parametric approaches can result in highly inefficient estimates and inflated Type I error rates (Huitema et al., 1996). Another common attribute of designs involving repeated measures is missing data. Failing to properly account for missing data limits the generalizability of the results, threatens internal validity, and can lead to biased and inefficient estimates (Hughes et al., 2019; Little & Rubin, 2002; Peng & Chen, 2018; Rubin, 1987).

Most statistical methods for SCEDs were developed for evaluating univariate (e.g., autoregressive integrated moving averages (ARIMA) models; Box & Jenkins, 1970) and bivariate relationships (e.g., simulation modeling analysis; Borckardt et al., 2008; standardized mean difference; Borenstein, 2009; percentage of non-overlapping data; Schlosser et al., 2008). However, the approaches mentioned thus far do not provide a method for examining the mechanism through which the intervention achieves its effects for a particular client. Recent advances in statistical methods for SCEDs have focused on adapting methods for mediation analysis to the SCEDs setting, and at least three methods have been proposed (Gaynor & Harris, 2008; Geuke et al., 2019; Miočević et al., 2020). While Gaynor and Harris (2008) relied on visual analysis to assess mediation, Geuke et al. (2019) presented the joint significance test to evaluate the significance of the indirect effect. Miočević et al. (2020) were the first to introduce a method for obtaining numerical estimates and credibility intervals for the indirect effect in SCEDs.

This paper aims to examine parameter estimation in a piecewise regression model by evaluating the statistical properties of the indirect effect (1) using three different methods for handling autocorrelation in repeated measures data and (2) examining the performance of multiple imputation in the presence of data that are missing completely at random (MCAR; Rubin, 1976). The following sections describe mediation analysis for a single mediator model, and we provide more details about the methods for estimating indirect effects in piecewise regression analysis proposed by Miočević and coauthors (2020).

Single Mediator Model

Mediation analysis is used to evaluate whether a variable acts as a mediator (M) transmitting the effect from the independent variable (X) to a dependent variable (Y). The effects of interest in the single mediator model can be estimated using three regression equations.

$$Y = i_1 + cX + e_1 \tag{2}$$

$$Y = i_2 + c'X + bM + e_2 \tag{3}$$

$$M = i_3 + aX + e_3 \tag{4}$$

In Equations (2)–(4), c represents the total effect of X on Y, c' is the effect of X on Y adjusted for the effect of the mediator M (also called the direct effect), b quantifies the relation between the mediator (M) and the dependent variable (Y) controlling for the effect of the independent variable (X), and a captures the relationship between X and M. The terms i_1, i_2, and i_3 represent intercepts, and it is assumed that the three error terms, e_1, e_2, and e_3, are uncorrelated and follow normal distributions with means of zero and variances σ^2_{e1}, σ^2_{e2}, and σ^2_{e3} (respectively).

The indirect effect can be computed as either the product of coefficients ab or as the difference of coefficients c-c', and the two approaches are equivalent in linear models with no missing values (Mackinnon et al., 1995). The significance of the indirect effect is generally evaluated using confidence or credibility intervals. Due to the asymmetry in the distribution of the product of two normal variates (i.e., ab; Craig, 1936; Lomnicki, 1967; Springer & Thompson, 1966), methods that use critical values from the distribution of the product and methods that make no distributional assumptions like the bootstrap lead to confidence intervals with the highest power (MacKinnon et al., 2007; MacKinnon et al., 2002; Miočević et al., 2017; Tofighi & MacKinnon, 2011; Yuan & MacKinnon, 2009). Subsequent sections describe methods for estimating indirect effects using piecewise regression analysis in combination with several techniques to model serially correlated error terms.

Estimating Indirect Effects in SCEDs using Piecewise Regression Analysis

In the single mediator model for SCEDs, both the mediator and outcome variables are repeatedly measured across at least two phases (i.e., the baseline phase and the intervention phase). We selected piecewise regression analysis because it allows for quantifying the change in the mediator as a result of the change in phase (a in Equation (4)) and change in outcome as a result of the change in the mediator (b in Equation (3)) controlling for the effect of phase.

Effects of interest for a single mediator model using piecewise regression analysis can be estimated using two equations (Miočević et al., 2020).[1]

$$\begin{aligned} M_t &= b_{0M} + b_{1M}\ time_t + b_{2M}\ phase_t \\ &\quad + b_{3M} + phase\_time_t + e_{M,t} \end{aligned} \tag{5}$$

$$\begin{aligned} Y_t &= b_{0Y} + b_{1Y}\ time_t + b_{2Y}\ phase_t \\ &\quad + b_{3Y}\ phase\_time_t + b_{4Y} M_t + e_{Y,t} \end{aligned} \tag{6}$$

As a result of the specific coding of the predictors, regression coefficients from the piecewise regression analysis provide estimates of the level of the first time point of phase A for the mediator (b_{0M}) and for the outcome (b_{0Y}), the trend in phase A for the mediator (b_{1M}) and for the outcome (b_{1Y}), the change in level at the start of phase B for the mediator (b_{2M}) and for the outcome (b_{2Y}), and the change in trend between the two phases for the mediator (b_{3M}) and for the outcome (b_{3Y}). We also obtain an estimate for the effect of the mediator on the outcome at time t (b_{4Y}). In their paper, Miočević and colleagues (2020) estimated the two equations in the Bayesian framework using OLS estimates of the regression coefficients as mean hyperparameters of the normal priors for intercepts and regression coefficients in Equations (5) and (6). In this paper, we opt for frequentist estimation instead.

There are two effects of phase on the mediator (a path in Equation (4)) in this context: the change in level (b_{2M}) and the change in trend (b_{3M}) between the two phases. If we define the a path as the change in level between the two phases, the indirect effect (the product ab; see Equations (3) and (4) in the section on the Single Mediator Model) of the phase change on the outcome can be quantified through the change in the level of the mediator. If we define the a path as the change in trend between the two phases, the indirect effect of the phase change on the outcome can be quantified through the change in the trend of the mediator. The effect of the mediator on the outcome (b in Equation (3)) is represented by b_{4Y} in Equation (6), and the direct effects (c' in Equation (3)) of phase on the outcome controlling for the mediated effect is decomposed into b_{2Y} (for the changes in level) and b_{3Y} (for the changes trend).

There are two indirect effects of interest in the piecewise regression model for SCEDs: (1) the product of coefficients $b_{2M}\, b_{4Y}$, representing the change in outcome variable due to the change in the level of the mediator following a change in phase, and (2) the product of coefficients $b_{3M}\, b_{4Y}$, representing the change in the outcome variable due to the change in the trend (slope) of the mediator following a change in phase.

Autocorrelation Modeling Techniques

In their review of 809 single-case designs, Shadish and Sullivan (2011) found that autocorrelation ranged from −0.931 to 0.736, and they noted that autocorrelation (ρ) is commonly underestimated when the number of observations is small. Using a procedure to correct for negative bias, the authors found that the mean value of autocorrelation in SCEDs was estimated to equal 0.752 in the seven AB designs evaluated and 0.320 in the 64 multiple-baseline designs examined (Shadish & Sullivan, 2011). In subsequent paragraphs, we describe procedures for modeling serial dependency among error terms.

In our first approach, we estimate the regression coefficients via ordinary least squares (OLS) regression. OLS regression is a consistent estimator even in the presence of serially correlated error terms; however, OLS is no longer the best linear unbiased estimator and does not yield correct standard errors in the presence of serially correlated error terms (Davidson & MacKinnon, 2003). Therefore, we use heteroskedasticity- and autocorrelation-consistent standard errors as proposed by Newey and West (Newey & West, 1987). In this approach, the exact form of serial correlation in the error terms does not need to be specified, and the procedure also allows for heteroscedasticity. Newey-West standard errors are available in most standard software packages, for example, in the sandwich package in R (Zeileis, 2004).

In our second approach, we explicitly model the serial correlation in the error terms. For this purpose, we assume that the error terms follow an AR(1) process. Thus, we add the following equations to Equations (5) and (6)

$$e_{M,t} = \rho_M e_{M,t-1} + v_{M,t} \tag{7}$$

$$e_{Y,t} = \rho_Y e_{Y,t-1} + v_{Y,t} \tag{8}$$

The autocorrelation coefficients ρ_M and ρ_Y quantify the strength of the serial dependency and are assumed to range between -1 and $+1$. The error terms $v_{M,t}$ and $v_{Y,t}$ in Equations (7) and (8) are assumed to be white noise and mutually independent. This class of models is well understood (Cochrane & Orcutt, 1949; Prais & Winsten, 1954) and can be estimated using generalized least squares (GLS) estimation. The GLS estimator is the best linear unbiased estimator (Davidson & MacKinnon, 2003). However, in practice, the true population values ρ_M and ρ_Y of the autocorrelation coefficients are unknown and need to be estimated from the data along with the regression coefficients. This procedure is known as feasible GLS (FGLS) and yields a non-linear estimator that is no longer unbiased. Furthermore, the small sample properties of FGLS are not known analytically. However, feasible GLS is asymptotically efficient (Davidson & MacKinnon, 2003). FGLS is implemented in several software packages, for example, in the orcutt package in R (Stefano et al., 2018). Note that the computation of the FGLS estimator in the above setting can be implemented using the so-called iterative Cochrane-Orcutt procedure (Stefano et al., 2018) which tends to outperform alternative two-step procedures for computing FGLS in small samples (Verbeek, 2017).

Our third approach is based on the same regression Equations (5) and (6) combined with the AR(1) Equations (7) and (8) for the error terms. In other words, we make the same modeling assumptions as in the case of FGLS. However, we use a different estimation technique, where the exact likelihood is computed via a state-space representation of the AR(1) process, and estimates are computed by a Kalman filter. This procedure is implemented in the stats package in R (R Core Team, 2020). The advantage of this procedure over FGLS lies in the possibility of including more complex patterns of serial correlation in the error term equations (e.g., moving average components, non-stationary integrated error terms). Throughout this paper, we focus on AR(1) error terms and thus expect that the results will be similar to those obtained by FGLS. We refer to the three approaches described above as NW, FGLS, and AR(1) throughout the remainder of the paper. We compare these three approaches to a standard OLS procedure that ignores the presence of autocorrelation.

Missing Data Handling Techniques

Missing data in SCEDs is common due to repeated measures of participants over time, resulting in noncompliance and participant attrition (Smith, 2012). There are three ways to categorize missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR is characterized by missing data that does not depend on observed data nor on the missing data, for example, when a random subset of participants' self-report data is lost. MAR, on the other hand, is a function of the observed data but not a function of the missing data. In a study evaluating confidence among university-aged men and women, for example, women may feel uncomfortable when asked to rate their appearance and choose not to answer questions related to physical appearance. In this case, the participant's gender results in nonresponse. Finally, MNAR occurs when the missingness is related to the unobserved data. When individuals with the lowest education are missing from a study evaluating educational outcomes, the missing data mechanism is MNAR. Improper handling of missing data using traditional methods such as listwise deletion and mean substitution can lead to loss of information, biased estimates, inefficiency, and introduce effects that are not supported by data (Little & Rubin, 2002). Modern approaches to handling missing data such as the expectation-maximization (EM) algorithm (developed by Dempster et al., 1977) and multiple imputation (MI) (Schafer & Graham, 2002) have gained favor over more traditional methods. Numerous studies have advocated for using maximum likelihood and EM methods for handling missing data in group multivariate designs (e.g., Horton & Kleinman, 2007; Ibrahim et al., 2005; Raghunathan, 2004). Velicer and Colby (2005a, 2005b) found that maximum likelihood is an effective strategy for handling missing data compared to listwise deletion, mean substitution, and mean of adjacent observations in time series data. In the subsequent paragraph, we describe MI in greater detail.

We examine MI proposed by Rubin (1987) for data that are MCAR in a simulation study. An imputation refers to one set of plausible values (m) for a missing observation, while MI represents multiple sets of plausible values ($m > 1$). When using MI, the missing value is replaced by a random sample of plausible values resulting in m complete datasets. The statistical analysis of interest (e.g., OLS regression) is then conducted on each m complete dataset separately. Finally, a single MI estimate and its standard error (SE) are estimated by combining results obtained from each m analysis using Rubin's rules (Rubin, 1987). Suppose $\widehat{Q}$ represents the estimate of a parameter Q (e.g., a regression coefficient) from the jth imputed data set. The pooled estimate is given by Equation (9).

$$\overline{Q} = m^{-1} \sum_{j=1}^{m} \widehat{Q}_j \tag{9}$$

The total variance of $\overline{Q}$, represented by T in Equation (12), is the weighted sum of the average within-imputation variance $\overline{U}$ (Equation (10)) and the between-imputation variance B (Equation (11)). The overall SE of $\overline{Q}$ is equal to the square root of T.

$$\overline{U} = m^{-1} \sum_{j=1}^{m} \widehat{U}_j \tag{10}$$

$$B = (m-1)^{-1} \sum_{j=1}^{m} \left[\widehat{Q}_j - \overline{Q}\right]^2 \tag{11}$$

$$T = \overline{U} + (m+1)^{-1} B \tag{12}$$

There are two methods for conducting multivariate MI in which values are missing on multiple variables: multivariate normal imputation (MVNI) and MI by chained equations (MICE). MVNI assumes that the incomplete variables follow a multivariate normal distribution (Lee & Carlin, 2010). MICE generates separate univariate imputation models for each variable with missing data (White et al., 2011). In the present study, we evaluate MICE as a missing data handling method in order to examine how one of the most commonly used R packages for handling missing data performs when adapted to piecewise regression analysis for SCEDs (van Buuren & Groothuis-Oudshoorn, 2011). To our knowledge, this is the first study that analyzes missing data in SCEDs in the presence of autocorrelated errors. Therefore, the aim of our simulation study is to examine what would happen if researchers just continued with their standard practice in the presence of autocorrelated errors.

Missing Data Handling in SCEDs

A review of missing data in SCED studies published by Chen et al. (2020) indicated that approximately 18% of studies (33 out of 182) contained missing data with a range of 1%–45% missing values. In general, studies reported a higher average percentage of missing values in the intervention phase (15%) compared to the baseline phase (6%) (Chen et al., 2020). Previous studies have investigated the performance of various missing data handling techniques in SCEDs (Smith et al., 2012; Peng & Chen, 2018; Chen et al., 2020; De, Michiel, Tanious, & Onghena, 2020). In a Monte Carlo simulation study, Smith et al. (2012) evaluated the performance of the EM procedure in terms of statistical power for data simulated as MCAR. Effect sizes were quantified using the standardized mean difference (Glass's Δ). They concluded that EM is effective at handling missing data across various levels of missingness (10%, 20%, 30%, and 40%) and lag-1 autocorrelation (0, 0.2, 0.4, 0.6), except when autocorrelation is large (i.e., 0.8). Peng and Chen (2018) applied MI to missing data from a published single-case ABAB design and examined effect sizes using Tau-U. They concluded that there are several advantages to MI over ad hoc methods such as mean substitution in that it avoids potential bias that can arise from omitting participants from an analysis, takes into account the uncertainty surrounding the imputed scores, and retains the design structure of the study. Chen et al. (2020) extended the findings from Smith et al. (2012) by examining the performance of EM in terms of relative bias (RB), root-mean squared error (RMSE), and relative bias of the estimated standard error (RBESE). They estimated the baseline slope, level shift, and slope change from a piecewise regression model for data simulated under a MAR mechanism for an AB design. The authors concluded that EM is an effective strategy for missing data handling in piecewise regression analysis for SCEDs. De et al. (2020) assessed the performance of three missing data handling methods for data simulated under MCAR: (1) randomized-marker method, (2) single imputation (SI) using an autoregressive integrated moving average (ARIMA) model, and (3) MI using multivariate imputation by chained equations (MICE). De et al. (2020) computed the mean difference (MD) and nonoverlap of all pairs (NAD) as their indicators of an intervention effect. The authors concluded that the randomized-marker method is a promising missing data handling technique as it outperformed the other methods in terms of statistical power while ensuring a low Type I error rate. Only one study (i.e., Chen et al., 2020) examined the performance of missing data handling methods for piecewise regression analysis in SCEDs. In this study, we aim to determine how MI performs for piecewise regression in SCEDs when data is MCAR. We simulated data under MCAR to reflect one possible scenario in practice, namely that missing data are due to the participant not filling out the questionnaire for a given measurement occasion due to a random event that prevented them from providing data for that observation. This results in complete data on the variables time, phase, phase_time in Equations (5) and (6) because those are part of the study design, whereas if the participant fails to complete the questionnaire, data are missing on the mediator and outcome at the same measurement occasion.

Table 1. Results from the Single Mediator Model Empirical Example.

	Autocorrelation Handling Method			
Parameter Estimate	OLS	NW	FGLS	AR(1)
b_{2M} b_{4Y}	−311.78	−311.78	56.57	56.96
b_{3M} b_{4Y}	−11.73	−11.73	1.82	1.83
95% CI for b_{2M} b_{4Y}	[−1320.40, 676.90]	[−901.03, 269.01]	[−40.21, 212.22]	[−38.72, 214.61]
95% CI for b_{3M} b_{4Y}	[−32.75, 8.74]	[−24.63, 0.89]	[−0.51, 5.63]	[−0.59, 5.74]
SE of b_{2M} b_{4Y}	507.50	297.79	64.37	64.86
SE of b_{3M} b_{4Y}	10.54	6.49	1.60	1.64

Note. Point and interval estimates of the indirect effect through the change in level and change in trend are displayed for OLS estimation and three methods for handling autocorrelation. b_{2M} b_{4Y} represents the indirect effect through the change in level. b_{3M} b_{4Y} represents the indirect effect through the change in trend. CI = confidence interval. SE = standard error.

Methods

Empirical Example

To illustrate the three approaches to modeling autocorrelation described above and compare them to OLS estimation, we apply the methods to an example data set from an AB SCEDs study. The study evaluated the effectiveness of a walking intervention for osteoarthritis in four individuals (O'Brien et al., 2016). Over 12 weeks, diary measures were taken twice daily on symptoms related to impairment (i.e., pain, pain on movement, and joint stiffness), cognitions (i.e., intentions, self-efficacy, and perceived controllability), and walking behavior (i.e., number of steps). The goal of the study was to evaluate the role of cognition in predicting outcomes in individually-tailored walking interventions for osteoarthritis. Cognitions, such as intention and self-efficacy, were hypothesized to transmit the effect of impairment on physical activity. The baseline phase was designed to obtain baseline measures and identify the cognitions that impacted walking activity in the participants. During the intervention phase, participants received an intervention that targeted the cognitions that were shown to strongly correlate with walking behavior.

We illustrate the proposed methods using data from a single participant, participant A. The number of observations was relatively even across phases, with 81 measurement occasions in the baseline phase and 89 measurement occasions in the intervention phase. For the single mediator model, the mediated effect of phase (X) on walking (Y), measured as the step count, through the intention to walk (M) was considered. The percentage of missing data in the study was equal to 2.37% and 3.55% for M and Y, respectively. Missing values on M and Y were imputed using the R package MICE (van Buuren & Groothuis-Oudshoorn, 2011). Next, an AR(1) model was fit to the residuals to inspect the autocorrelation. Estimates for the serial correlation among the residuals were −0.54 and −0.81 for M and Y, respectively. Parameter estimates were obtained using piecewise regression analysis, and autocorrelation was modeled according to the three proposed approaches which were compared to OLS. The analyses were conducted in R (R Core Team, 2020) using the package RMediation to compute the Monte Carlo confidence intervals (CI) for the mediated effect (Tofighi & MacKinnon, 2011).[2] The annotated R syntax for the analyses is provided in the Supplemental Materials available at https://osf.io/ahpkm/?view_only=d9144ffbc4bc4af28df31b03164ed6b2.

Point and interval estimates are displayed in Table 1. The interval estimates for the indirect effect through the change in level for the single mediator model were consistent across methods with the significance test indicating that the 95% CIs contained zero. The interval estimates for the indirect effect through the change in trend for the four approaches to handling autocorrelation also led to the same conclusion about the significance test in that the 95% CIs contained zero. Nevertheless, there were noticeable differences in the point and interval estimates across methods. The point estimate for the indirect effect through changes in level and trend were closer to zero for FGLS and AR(1) than OLS and NW. The interval widths for both the indirect effect through changes in level and trend were substantially smaller for FGLS and AR(1) than for OLS and NW, indicating more precision in the estimates. The standard error of the point estimate for the indirect effect for changes in level and trend was largest for OLS, followed by NW. AR(1) and FGLS had smaller standard errors than both OLS and NW.

Simulation Studies

Simulation studies were performed to investigate the number of time points required to attain acceptable statistical properties for point and interval estimates of the indirect effect in a single mediator model. We assessed the bias, relative bias, efficiency (the standard deviation of the point estimate across replications), power, Type I error rate, coverage, and interval width. Bias and relative bias were used to assess the accuracy of the point estimate of the indirect effect. Since bias is affected by the size of the indirect effect, relative bias is the preferred measure of accuracy. Relative bias is computed as the difference between the value of the indirect effect in the population and the estimate of the indirect effect divided by the value of the indirect effect in the population. When the true indirect effect is zero, relative bias is undefined. Values of relative bias between −0.10 and 0.10 were considered acceptable (Kaplan, 1988). The standard deviation of the estimate of the indirect effect over replications was a measure of efficiency, where higher standard deviation values indicated lower efficiency. Power was defined as the proportion of confidence intervals for the indirect effect that excluded zero when the indirect effect is nonzero. Values of 0.8 and higher were considered desirable. Type I error rate was computed as the proportion of confidence intervals that excluded zero when the true value of the indirect effect was zero. A Type I error rate of 0.05 was deemed desirable, and values between 0.025 and 0.075 were acceptable (Bradley, 1978). Coverage was defined as the percentage of confidence intervals that contained the true value of the indirect effect, and values of coverage between 0.925 and 0.975 were considered close to the nominal level of 0.95 according to Bradley's robustness criterion (1978). Interval width was defined as the difference between the upper confidence limit and the lower confidence limit. Lower interval widths represent higher precision. The R code for the simulation studies is available in the Supplemental Materials.

Single Mediator Model: No Missing Data

A total of 1000 replications were simulated from piecewise regression models using Equations (5) and (6). M and Y were simulated as continuous variables from an AB design. Ferron et al. (2010) reported that the median number of observations in SCEDs was 24 with a range of seven to 58. To simulate realistic conditions, we evaluated sample sizes (N) of 20 and 30. We also chose to consider sample sizes larger ($N = 60$ and 100) than those typically observed in SCEDs. Although the baseline phase is typically shorter than the treatment phase in SCEDs (Shadish & Sullivan, 2011), we opted for an equal number of observations in the baseline and treatment phases to isolate the effect of autocorrelation on the performance of our methods. When N = 60, for example, we assigned 30 observations to the baseline phase and 30 observations to the intervention phase. Values of 0 and 2 were simulated for the a path defined as the change in level (b_{2M}; a_level), and values of 0 and 0.2 were simulated for the a path defined as the change in trend between the two phases (b_{3M}; a_trend). The nonzero effect sizes for the a paths come from the conventions in the simulation literature on SCEDs (the a_level of 2 and a_trend of 0.2 were used in, e.g., Moeyaert et al., (2013a) and Moeyaert et al., (2013b)). Values of 0 and 0.59 were chosen for the b path (b_{4Y}). The nonzero b path value stems from the conventions in the simulation literature on mediation models (the value of 0.59 is considered as a large value for the b path in the single mediator model in, e.g., Fritz & MacKinnon (2007); MacKinnon et al,. (2004), and Wang and Preacher (2015)). We included effect sizes of 0 for the a and b paths to evaluate the Type I error rates of the methods.[3] Four different values of ρ (0, 0.1, 0.5, and 0.9) were simulated, and we analyzed the simulated data using four methods: OLS, NW, FGLS, and AR(1). Statistical properties were evaluated for 128 (4 × 2 × 2 × 2 × 4) combinations of parameters.

The simulation study was carried out in R (R Core Team, 2020). To generate data for M and Y, first, a deterministic portion of M and Y was simulated based on parameter values for a given condition. Four levels of autocorrelation (ρ) were simulated (no AR, AR.small, AR.medium, AR.large). In the no AR(1) condition, normal residuals with means of 0 and standard deviations of 1 were added to the deterministic portion using the rnorm command. In the conditions where low, medium, and large autoregressive effects were evaluated, residuals were added to the deterministic function using the arima.sim() function based on the specified value of ρ (0.1, 0.5, or 0.9). For OLS estimation, estimates of the parameters in Equations (5) and (6) were obtained using the lm() function. Using NW standard errors, regression coefficients were identical to those obtained from OLS estimation. The vcovHAC() function in the R package sandwich (Zeileis, 2004) was used to estimate a HAC covariance matrix, and coeftest() in the R package lmtest (Zeileis & Hothorn, 2002) was used to obtain the estimates for the HAC standard errors. The coefficients and standard errors for FGLS estimation were obtained using the cochrane.orcutt() function in the orcutt package in R (Stefano et al., 2018). Finally, in order to fit an AR(1) model, the data were transformed into a time series object using ts(). The arima() function was used to fit a model to the time series data with an autoregressive structure of order one. Point estimates of the indirect effect ab using OLS and our three autocorrelation handling methods were computed through the change in level and the change in trend. The RMediation package (Tofighi & MacKinnon, 2011) was used to compute 95% confidence intervals using the Monte Carlo method with the medci() function (Mackinnon et al., 2004). Finally, statistical properties of the point and interval estimates were computed as described in the previous paragraph for each iteration of the simulation study.

Single Mediator Model: Missing Data

The simulation for the single mediator model with missing data on M and Y was performed in a similar fashion to the simulation for the single mediator model without missing data. We considered the same parameter values for b_{2M}, b_{3M}, and b_{4Y}, sample sizes, autoregressive values, and autoregressive handling methods as in the single mediator model without missing data. Two proportions of missing data (20% and 50%) were simulated under an MCAR condition. Statistical properties were evaluated for 256 (4 × 2 × 2 × 2 × 4 × 2) combinations of parameters.

The simulation study was carried out in R (R Core Team, 2020). The first step in data generation for M and Y was the same as in the complete case scenario. Following the simulation of N values of M and Y with a specific value of ρ (0, 0.1, 0.5, or 0.9), we used the ampute() function from the R package

MICE (van Buuren & Groothuis-Oudshoorn, 2011) to remove a specified proportion of observations (20% or 50%) in both M and Y under MCAR. The mice() function was then used to impute missing values on M and Y. Following the recommendation of White et al. (2011), who proposed that the number of imputations should be at least equal to the proportion of missing data (e.g., 30% missing data requires at least 30 imputations), we requested 100 imputations, and we chose five iterations according to the recommendation of van Buuren and Groothuis-Oudshoorn (2011). OLS regression was conducted on the 100 complete data sets using the with() function, and the estimates for the regression coefficients were combined into one estimate using the pool() function. Estimates for the NW standard errors, as well as the FGLS and AR(1) estimates, were obtained using the same procedures as in the simulation study with complete cases for each imputed data set, followed by the pooling together of the estimates using the pool() function. Point and interval estimates of the indirect effect ab were computed following the same procedures as in the simulation study with complete cases. Finally, statistical properties of the point and interval estimates were computed for each iteration of the simulation study.

Results

Single Mediator Model: No Missing Data

Bias and Efficiency. Through the change in level, the point estimates for the indirect effect were unbiased (Figure S1 found in Supplemental Materials). The range of relative bias generally increased as the autoregressive effect increased for all autocorrelation handling methods. When $N = 60$, the four methods performed comparably in terms of relative bias across all levels of autocorrelation and parameter combinations. When $N = 20$, 30, and 100 and $\rho = 0.9$, the mean relative bias over 1000 replications was greater for OLS and NW than FGLS and AR(1). FGLS followed by AR(1) resulted in the lowest mean relative bias when the amount of simulated autocorrelation was high. In terms of efficiency, at all values of N, the methods performed similarly for $\rho = 0$, 0.1, and 0.5. When $\rho = 0.9$, the standard deviation across replications was greater for OLS and NW than FGLS and AR(1) which performed similarly. As the sample size increased, the standard deviation over replications generally decreased for $\rho = 0$, 0.1, and 0.5, and this effect was most noticeable for nonzero b paths (Figure S2A). The mean standard deviation increased as the sample size increased when $\rho = 0.9$ for OLS and NW.

The estimates of the indirect effect through changes in trend were unbiased in the majority of conditions. However, when $N = 20$ and $\rho = 0.9$, the relative bias across replications was greater than 0.10 for all autocorrelation handling methods, where OLS and NW had a higher mean relative bias than AR(1) and FGLS (Figure S1). The range of relative bias generally increased as the autocorrelation increased for all methods. As the sample size increased, the standard deviation of the point estimate decreased across all autocorrelation handling techniques, and this effect was most noticeable at $\rho = 0.9$ for nonzero b paths (Figure S2). The methods performed similarly for $\rho = 0$, 0.1, and 0.5 in terms of efficiency. At $\rho = 0.9$, AR(1) methods had lower standard deviation values than FGLS, NW, and OLS in most conditions, although the differences between methods were less pronounced than through changes in level. Lower values of standard deviation for AR(1) were most noticeable when $b = 0$.

Power

Through the change in level, power increased as the sample size increased for most parameter combinations (Figure 1). When $N = 20$ and 30, power was below 0.8 for all autocorrelation handling methods and at all levels of simulated autocorrelation. When $N = 60$ and $N = 100$ and $\rho = 0$ and 0.1, power exceeded the nominal value of 0.8. When $N = 60$ and $\rho = 0.5$, power was equal to 0.8 for OLS and below 0.8 for NW, FGLS, and AR(1). Power was unacceptable for all methods when $\rho = 0.9$ and $N = 60$ and 100. OLS had the highest values of power when $\rho = 0.9$ at large sample sizes, followed by FGLS and AR(1) which performed comparably. NW yielded the lowest power.

Through the change in trend, power increased as the sample size increased. Power was below 0.8 for small sample sizes (i.e., $N = 20$ and 30) (Figure 1). When $N = 60$ and 100 and $\rho = 0$, 0.1, and 0.5, power was above 0.8. When $\rho = 0.9$ and $N = 60$ and 100, power was below 0.8 for NW, FGLS, and AR(1). OLS yielded acceptable power at $N = 100$. FGLS had the lowest power of all the methods at $\rho = 0.9$.

Type I Error Rate

Through the change in level, Type I error rates generally increased as the sample size increased for OLS and NW, while Type I error rates decreased or remained stable for FGLS and AR(1) at large autoregressive effects (Figure 2(a)). Type I error rates equal to or below 0.075 were observed for FGLS and AR(1) in the majority of parameter combinations when $b = 0$. When $b = 0$ and $\rho = 0.9$, OLS and NW interval estimates had Type I error rates above 0.075. When $b = 0.59$ and $\rho = 0.5$ and 0.9, Type I error rates above 0.075 were observed for all methods. OLS and NW had considerably higher Type I error rates than FGLS and AR(1) at $N = 60$ and 100 and when the autoregressive effect was medium or large.

Through the change in trend, Type I error rates generally increased for NW and OLS and increased or remained at the same level for FGLS and AR(1) as the sample size increased at large autoregressive effects (Figure 2(b)). Type I error rates equal to or below 0.075 were observed for FGLS and AR(1) for all parameter combinations when $b = 0$. When $b = 0$ and $\rho = 0.9$, OLS and NW interval estimates had Type I error rates above 0.075. When $b = 0.59$ and $\rho = 0.5$ and 0.9, all methods had excessive Type I error rates, whereas FGLS and AR(1) had lower Type I error rates than OLS and NW.

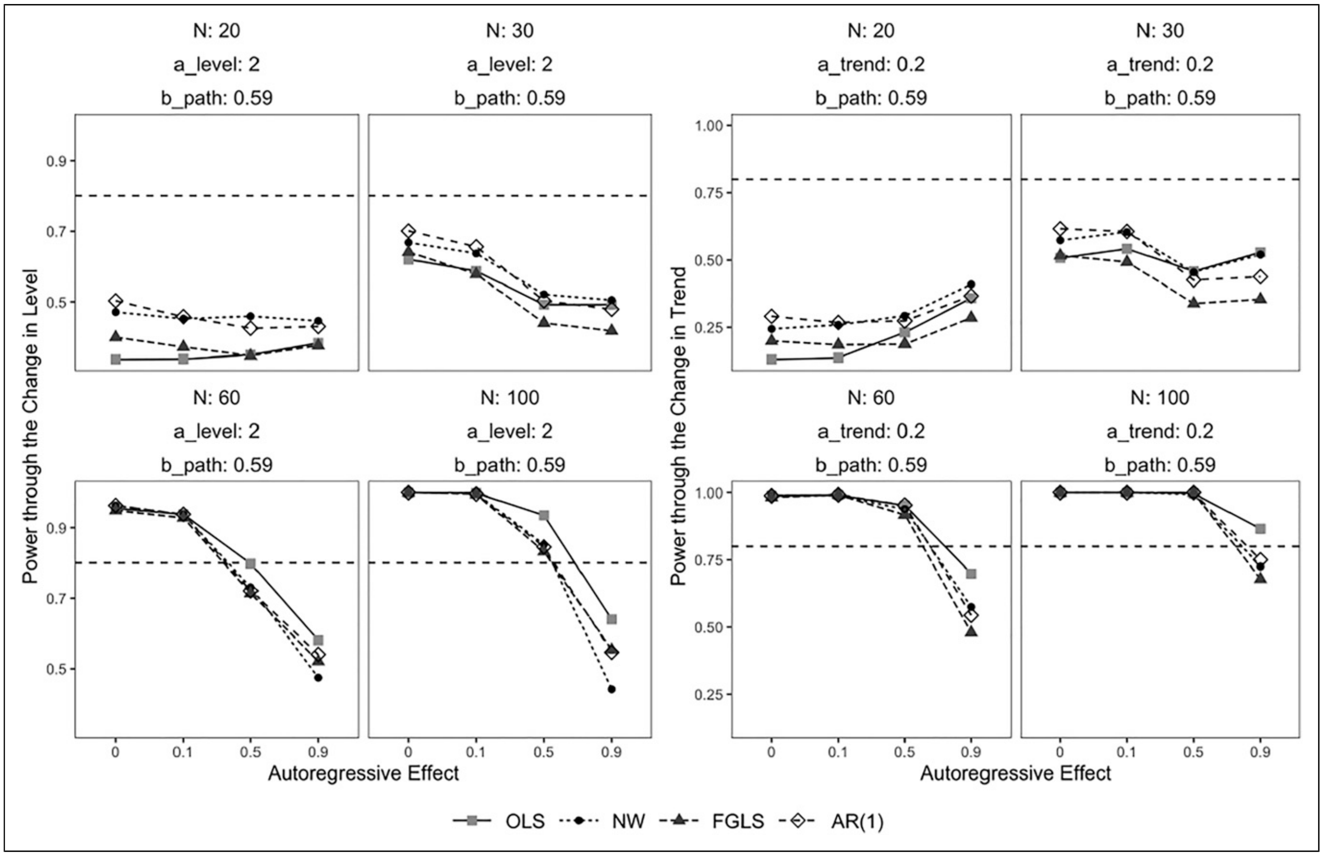

Figure 1. Power of the Interval Estimate of the Indirect Effect through the Change in Level and Trend
Note. Power of the interval estimate of the indirect effect defined through the change in level and trend over 1000 replications. The dotted line represents power of 0.8. Different values for the *a* path defined as the change in trend did not impact the power for the change in level, and different values for the *a* path defined as the change in level did not impact the power for the change in trend. a_level = *a* path as the change in level. a_trend = *a* path as the change in trend. b_path = *b* path. N = sample size.

Coverage

Through the change in level, coverage was consistent across sample sizes for FGLS and AR(1), while coverage decreased for NW and OLS as the sample size increased in the majority of conditions (Figure S3A). FGLS and AR(1) generally had coverage above 0.925 when $b = 0$. Coverage was below 0.925 for NW and OLS when $\rho = 0.9$, $b = 0$, and $N = 60$ and 100. When $\rho = 0.9$ and $b = 0.59$, coverage increased as the sample size increased for FGLS and AR(1), while coverage decreased for OLS and NW. For nonzero b paths and large autoregressive effects, coverage approached 0.925 for FGLS and AR(1), while coverage was well below 0.925 for OLS and NW at larger sample sizes.

Through the change in trend, coverage decreased as sample size increased for OLS and NW at high levels of simulated autocorrelation in the majority of conditions (Figure S3B). Coverage above 0.925 was obtained for FGLS and AR(1) when $b = 0$ at all values of autocorrelation and sample sizes. NW and OLS had coverage below 0.925 when $\rho = 0.9$, $b = 0$, and $N = 30$, 60, and 100. When $b = 0.59$ and $\rho = 0.9$, coverage was below 0.925 for all methods, whereas FGLS and AR(1) had higher coverage than NW and OLS.

Interval Width

Through the change in level, interval width was larger at $N = 20$ and 30 than at $N = 60$ and 100, and the performance across methods was consistent for smaller sample sizes (Figure S4A). There were discrepancies in the performance of methods at larger sample sizes and autoregressive effects. When $b = 0$, $\rho = 0.9$, and $N = 60$ and 100, FGLS and AR(1) had smaller interval widths than OLS and NW. When $b = 0.59$, $\rho = 0.9$, and $N = 60$ and 100, the interval estimates were more precise for OLS than the other methods.

Through the change in trend, interval width decreased as the sample size increased (Figure S4B). The methods performed similarly when $\rho = 0$, 0.1, and 0.5. When $b = 0.59$, $\rho = 0.9$, and $N = 20$ and 30, FGLS had larger interval widths than the other methods.

Single Mediator Model: Missing Data

Bias and Efficiency. Through the change in level, the point estimates of the indirect effect were unbiased when the proportion of missingness was 0.2 in most conditions (Figure S5A). When the proportion of missingness was large (0.5), the

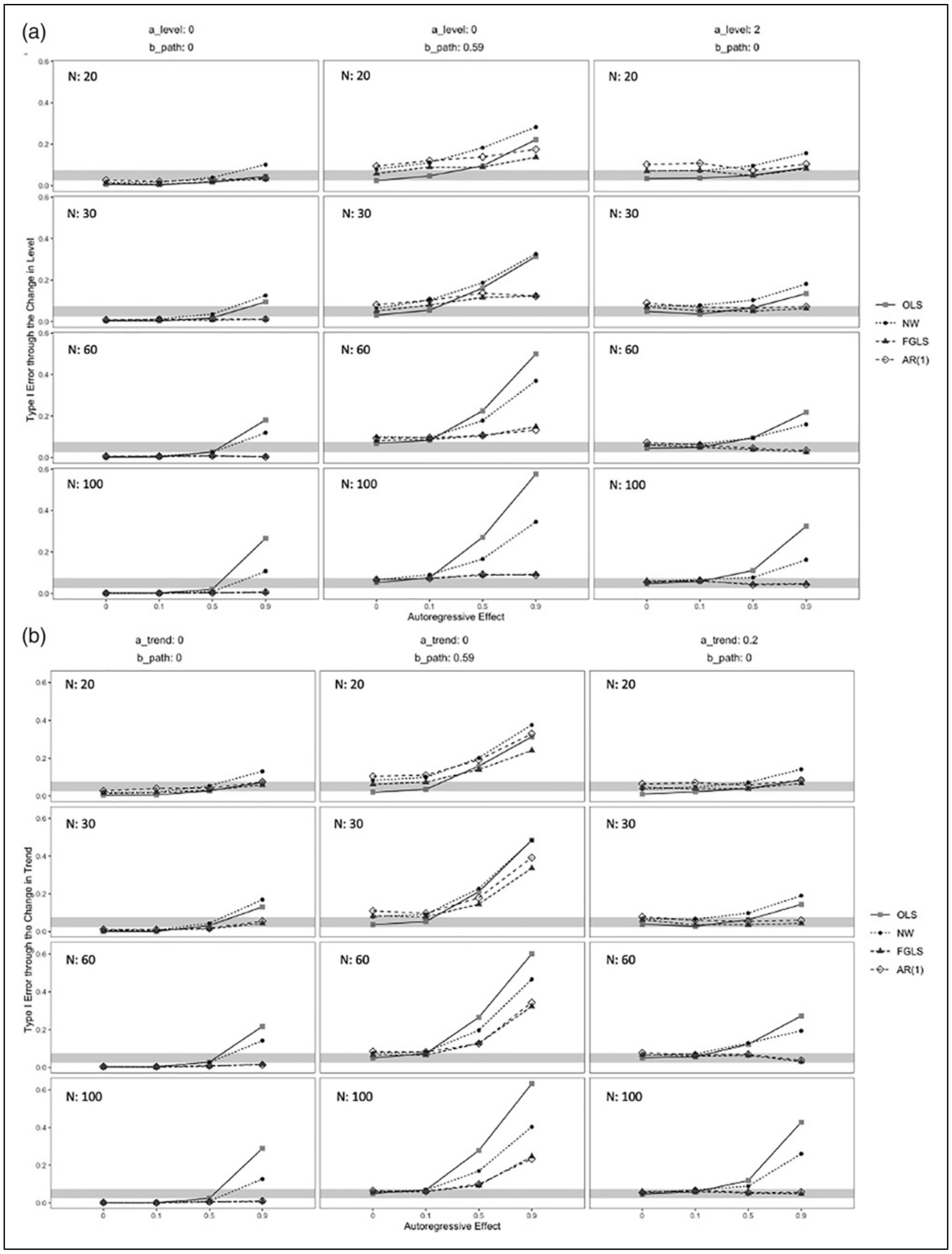

Figure 2. Type I Error of the Estimate of the Indirect Effect through the Change in Level and Trend.
Note. Type I error rates of the interval estimate of the indirect effect defined as the change in level and trend over 1000 replications. The shaded area represents the acceptable range of Type I error rates between 0.025 and 0.075. Different values for the *a* path defined as the change in trend did not impact the Type I error rates for the change in level, and different values for the *a* path defined as the change in level did not impact the Type I error rates for the change in trend. a_level = *a* path as the change in level. a_trend = *a* path as the change in trend. b_path = *b* path. N = sample size.

point estimates were biased for most combinations of parameter values. Relative bias generally decreased as the autoregressive effect increased at $N = 60$ and 100. The results were consistent across autocorrelation handling methods. The standard deviation of the point estimate generally increased as the autoregressive effect and sample size increased (Figure S6). There were no major differences across missingness proportions of 0.2 and 0.5. NW and OLS were less efficient estimators than FGLS and AR(1) in most conditions when the autoregressive effect was large, and this effect was most evident at $N = 60$ and 100.

Through the change in trend, the relative bias generally decreased as the sample size increased (Figure S5B). Relative bias was unacceptable when missing = 0.5 and $N = 20$ and 30 for all methods. When $N = 60$, the point estimates were unbiased regardless of the proportion of missing data. When

N = 100, the estimates were biased at several levels of autocorrelation. Relative bias was generally higher for larger proportions of missingness. The results were consistent across autocorrelation handling techniques. The standard deviation of the point estimate generally increased as the autoregressive effect increased and decreased as the sample size increased (Figure S7). The results were generally consistent across proportions of missingness and autocorrelation handling method. However, when missing = 0.2, b path = 0, and ρ = 0.9, FGLS and AR(1) had lower values of standard deviation than OLS and NW.

Power

Through the change in level, power decreased as the proportion of missing data increased, and power increased as the sample size increased (Figure S8A). Power was inadequate for all autocorrelation handling methods when N = 20 and 30 at both levels of missingness. When N = 60, missing = 0.2, and ρ = 0.5 and 0.9, power was below 0.8. When N = 60 and missing = 0.5, power was unacceptable at all levels of autocorrelation. When N = 100, missing = 0.2, ρ = 0.9, values of power were below 0.8. At N = 100 and a_trend = 0.2, power was below 0.8. When N = 100 and ρ = 0 and 0.1, power was above 0.8. Power was consistently higher for OLS in most conditions when ρ = 0.9 and N = 60 and 100.

Through the change in trend, power generally decreased as the proportion of missingness increased, and power increased as the sample size increased (Figure S8B). Power was inadequate for all methods when N = 20 and 30. Power was generally above 0.8 when missing = 0.2 and N = 60 and 100. When N = 60, missing = 0.2, and ρ = 0.9, power was below 0.8. When N = 100, missing = 0.2, and ρ = 0.9, power was above 0.8 for OLS and slightly below 0.8 for the other autocorrelation handling techniques. When missing = 0.5, power was below 0.8 in most conditions. However, when N = 100, missing = 0.5, and the a_level = 0, power was above 0.8 when ρ = 0.5 for all methods, and near 0.8 for ρ = 0, 0.1, and 0.9. Power was slightly higher for OLS than NW, FGLS, and AR(1) when ρ = 0.9 and N = 60 and 100 for all parameter combinations.

Type I Error Rate

Through the change in level, Type I error rates were within the acceptable range or below 0.025 when ρ = 0 and 0.1 across both levels of missingness and for all autocorrelation handling methods (Figure S9). As the sample size increased, the Type I error rates increased when ρ = 0.5 and 0.9 in several conditions. When ρ = 0.5, b path = 0.59, and missing = 0.2, the Type I error rates were above 0.075 when N = 30, 60, and 100. When ρ = 0.9 and N = 60 and 100, the Type I error rates were unacceptable in most conditions for OLS and NW and in several conditions for FGLS and AR(1).

Through the change in trend, Type I error rates were acceptable or below 0.025 when ρ = 0 and 0.1 at both levels of missingness (Figure S10). As the sample size increased, the Type I error rates increased when ρ = 0.5 and 0.9 in several conditions across both proportions of missingness. All methods had instances of Type I error rates above 0.075 when ρ = 0.5 and 0.9, N = 60 and 100, b path = 0.59, and missing = 0.2 and 0.5. OLS and NW performed worse than FGLS and AR(1) at large autoregressive effects for N = 30, 60, and 100 in several conditions, and this effect was most noticeable for nonzero b paths.

Coverage

Through the change in level, all autocorrelation handling techniques had values of coverage within or above the acceptable range when ρ = 0 and 0.1 (Figure S11). As the sample size increased, values of coverage generally decreased when ρ = 0.5 and 0.9. Coverage below 0.925 was observed in several parameter combinations when ρ = 0.5 and 0.9 and N = 30, 60, and 100, whereas OLS and NW had lower coverage than FGLS and AR(1) across both levels of missingness. A larger proportion of missingness resulted in higher coverage for OLS and NW when ρ = 0.9.

Through the change in trend, the methods had coverage within or above the robustness criterion for all parameter combinations when ρ = 0 and 0.1 (Figure S12). As the sample size increased, values of coverage generally decreased when ρ = 0.5 and 0.9. When missing = 0.2, b path = 0.59, ρ = 0.5 and 0.9, and N = 30, 60, and 100, coverage was below 0.925. OLS and NW performed worse than FGLS and AR(1).

Interval Width

Through the change in level, interval width generally increased as the autoregressive effect and proportion of missingness increased (Figure S13). As the sample size increased, the interval width decreased. OLS had slightly smaller interval widths than NW, FGLS, and AR(1) when the autoregressive effect was large at N = 60 and 100.

Through the change in trend, interval width generally increased as the autoregressive effect and the percentage of missing data increased (Figure S14). As the sample size increased, the interval width decreased. When ρ = 0.9 and N = 60 and 100, OLS had consistently smaller interval widths than NW, FGLS, and AR(1).

Discussion

In the present study, we evaluated various techniques for modeling serially correlated error terms and examined the performance of MI as a missing data handling technique for a MCAR mechanism for piecewise regression analysis in SCEDs. Specifically, we investigated the performance of the autocorrelation handling methods and MI in terms of the statistical properties of the point and interval estimates of the indirect effect for a single mediator model. After data were simulated with different values of autocorrelation (ρ = 0, 0.1, 0.5, and 0.9) and missing data (0% in the first simulation study and 20% and 50% in the second simulation study), the performance of the methods was assessed in terms of bias, efficiency, power, Type I error rate, coverage, and interval width.

Results from the single mediator model simulation without missing data revealed that OLS, NW, FGLS, and AR(1) generally have unbiased point estimates. The results were consistent across methods for small and medium autoregressive effects. At large autoregressive effects (i.e., ρ = 0.9), OLS and NW performed worse than FGLS and AR(1) in terms of relative bias for some parameter combinations. The four methods were less efficient at larger simulated autoregressive effects and smaller sample sizes. AR(1) and FGLS were more efficient estimators of the indirect effect at large autoregressive effects. Type I error rates were desirable or below 0.025 in most parameter combinations and sample sizes for AR(1) and FGLS. As sample size increased, however, Type I error rates increased for OLS and NW, particularly for large autoregressive effects. A similar pattern was observed in terms of coverage, where the performance of our methods was hindered by large autoregressive effects and sample sizes for OLS and NW. Coverage was acceptable or above 0.975 in most conditions for AR(1) and FGLS. The performance of NW and OLS in terms of Type I error rate and coverage were noticeably worse compared to AR(1) and FGLS at ρ = 0.9, and the discrepancy between performance increased as the sample size increased. When the sample size was small (i.e., N = 20 and 30), low power was achieved for all methods. At larger sample sizes (i.e., N = 60 and 100), satisfactory power was achieved in the majority of conditions except for when the simulated autocorrelation was substantial (ρ = 0.9). OLS had higher power than FGLS, AR(1), and NW in all conditions when ρ = 0.9 and N = 60 and 100.

Results from the simulation with missing data revealed that in general point estimates were unbiased when the proportion of missingness was 20%. However, the performance of the methods was unacceptable in terms of relative bias at several parameter combinations and sample sizes when a large proportion of missingness (50%) was introduced in the data. This finding is supported by Chen et al. (2020) who found a high missing rate negatively impacted the performance of EM in terms of relative bias. However, it is worth noting that the highest missing rate evaluated was 30%, and they simulated a lower missing rate for the A phase than the B phase (Chen et al., 2020). Our results indicate that the missing rate on its own affected the relative bias of the point estimate. At N = 60 and 100 and when the rate of missing data was equal to 20% and 50%, the relative bias generally decreased as the autoregressive effect increased. This finding is consistent with the literature demonstrating that the inclusion of auxiliary variables when the correlation among variables is high may be beneficial for the performance of multiple imputation models (Hardt et al., 2012). Thus, the inclusion of M and Y as auxiliary variables in our simulations may have improved the performance of our methods at large autoregressive effects. Chen et al. (2020) also found that relative bias was reduced for larger autoregressive effects in several conditions when the proportion of missingness in the intervention phase was equal to 20%. Consistent with the results from the simulation without missing data, the methods were less efficient at larger autoregressive effects, and the results were not negatively impacted by larger proportions of missingness. The results of Chen et al. (2020) are consistent with this finding, where the effect of the missing rate on the precision of the estimates was minimal compared to the impact of the autocorrelation. FGLS and AR(1) were more efficient estimators of the indirect effect when ρ = 0.9. Values of power below 0.8 were observed for all parameter combinations at N = 20 and 30. Power was acceptable at larger sample sizes excluding conditions when the percentage of missing data was equal to 50%. Type I error rates were acceptable for all methods when the level of simulated autocorrelation was small. However, when ρ = 0.5 and 0.9, Type I error rates were above 0.075 for larger sample sizes. Interestingly, Type I error rates tended to increase for medium and large autoregressive effects as the sample size increased. OLS had consistently higher Type I error rates at large autoregressive effects. Higher proportions of missingness did not negatively impact the Type I error rates. Coverage was within the acceptable range or above the upper limit of the nominal interval for small autoregressive effects. Consistent with the results of Type I error rates, coverage decreased as sample size increased for large autoregressive effects. Coverage was below 0.925 in several conditions when at ρ = 0.9. OLS and NW performed worse than FGLS and AR(1), and this effect was most noticeable through changes in level and at larger sample sizes. Coverage, like Type I error rates, decreased as the simulated autocorrelation increased. Interestingly, the performance of our methods improved in terms of the Type I error rate and coverage when a higher proportion of missingness was simulated. Finally, interval width increased as the percentage of missing data and autoregressive effect increased and decreased as the sample size increased.

The findings from our study revealed that (1) FGLS and AR(1) are promising methods for modeling autocorrelation, and (2) MI is a valuable missing data handling technique in piecewise regression analysis in SCEDs. The first major finding is supported by the low Type I error rates, high coverage, and high efficiency observed when low missing data rates (0% and 20%) were simulated. The second major finding is supported by the similar performance of the methods across missing data rates of 0% and 20% in terms of relative bias, efficiency, and power, and the superior performance of the methods in terms of Type I error and coverage for larger proportions of missingness when a large autoregressive effect was simulated. In light of these findings, we provide recommendations for applied researchers in subsequent paragraphs.

When acceptable Type I error rates, coverage, and high efficiency are sought out, AR(1) and FGLS would be recommended at all sample sizes, autoregressive effects, and proportion of missingness. However, the estimates were biased, and power was below 0.8 when a large proportion of missing data was simulated for all methods. When higher values of power are desired, the choice of method depends on the amount of autocorrelation in the data. OLS would be suggested when the amount of autocorrelation is medium or large. However, one should note that this choice of parameter estimation comes at the cost of an increased Type I error rate. When the autoregressive effect in the data is minimal, FGLS and AR(1) would be recommended. Procedures for estimating autocorrelation for the methods evaluated in the paper are provided in the Supplemental Materials. All autocorrelation handling methods resulted in values of power below 0.8 when the sample size was small. In order to achieve adequate power to detect indirect effects using the methods in this study, larger sample sizes (i.e., $N = 60$ and 100) are recommended. When $N = 60$ and 100, the power was near the acceptable value of 0.8 for all methods when ρ ranged from 0 to 0.5, and the percentage of missing data was equal to 0% and 20%.

Limitations and Suggestions for Future Research

There are several limitations to the current study. First, it may be unrealistic to expect researchers to collect the 60 to 100 data points per participant necessary to attain adequate power to detect indirect effects. As Shadish and Sullivan (2011) noted, 90.6% of single-case design studies had less than 50 observations. Fortunately, the development of smartphones, tablets, and handheld computers has revolutionized our ability to collect data. Advancements in real-time monitoring technology have facilitated the use of ecological momentary assessment (EMA) in which researchers acquire repeated data of participants' behaviors and experiences (Shiffman et al., 2008). EMA can readily document the behavior of an individual across time, revealing the effects of an intervention or treatment. New technologies have also promoted the use of passive real-time monitoring (Kleiman & Nock, 2017). Passive monitoring involves collecting data without requiring active participation and data entry from the individual. This enables researchers to collect data passively using features on smartphones such as screen time and social media activity (Vilardaga et al., 2014). The advantages of real-time monitoring technology in SCEDs are detailed thoroughly in Bentley et al. (2019).

Despite the finding that power increased as sample size increased when the proportion of missing data was large, power was below the nominal level, and relative bias exceeded 0.10 when 50% missing data was simulated. More methodological work is needed to develop optimal missing data handling methods to reduce bias and increase power for piecewise regression analysis in SCEDs. Several methods for multivariate data imputation have been proposed, including imputation based on maximum likelihood (MLMI; von Hippel & Bartlett, 2021) and predictive mean matching (PMM; Morris et al., 2014). Various R packages for performing imputation in time series data have been developed (Moritz et al., 2015), including the R package imputeTS (Moritz & Bartz-Beielstein, 2017) for univariate time series imputation and the R package Amelia II (Honaker et al., 2011), a bootstrap-based EM algorithm implemented for imputing missing values in multivariate time series data.

Another limitation of our study lies in the choice to consider only positive autocorrelations, yet negative autocorrelations have been reported in SCEDs (Harrington & Velicer, 2015; Parker et al., 2005). Studies have revealed differences in the performance of missing data handling methods under negative autocorrelations with time series data (Velicer & Colby, 2005a, 2005b). Future simulations should examine the performance of MI under both positive and negative autocorrelation values. The negative relationship between relative bias and autoregressive effect also warrants further investigation. Our simulation study and empirical example evaluated AB designs, and researchers may be interested in other types of SCEDs, such as multiple-baseline designs or alternating treatment designs. Shadish and Sullivan (2011) found that the multiple-baseline design is most commonly used in SCED research. However, methods for obtaining numerical estimates of indirect effects for multiple-baseline and intervention designs have yet to be described. Furthermore, we assessed a single mediator model, and often researchers are interested in evaluating more than one mediator. Future research is needed to identify optimal techniques for modeling autocorrelation and handling missing data for two mediator models. Additionally, we evaluated data that followed an MCAR mechanism, although data that is MAR, where a missing observation depends on the observed data, may be more realistic in empirical SCEDs.

Future research is needed to examine the effects of lagged and cross-lagged variables in piecewise regression models for SCEDs. The proposed method does not allow for lagged effects, for example, of the mediator M_t at a measurement occasion t to the outcome Y_{t+1} at the subsequent measurement occasion. Furthermore, we assume equidistant time intervals between measurement occasions. Mediation analysis with lagged effects can be done, for example, using multivariate time-series models (Lutkepohl, 2005), cross-lagged panel models (Usami et al., 2019; Zyphur et al., 2020), or non- and semi-parametric models for causal mediation analysis (Shpitser, 2013; Zheng & van der Laan, 2017). The assumption of equidistant time intervals can be relaxed by using continuous time models (Albert et al., 2019; Deboeck & Preacher, 2016).

In the present study, we did not distinguish between the number of time points in the baseline and intervention phase. However, it is common in SCEDs that the length of the treatment phase exceeds that of the baseline phase (Ferron et al., 2010; Shadish & Sullivan, 2011). Given that this is the first simulation study to examine methods for handling missing data and autocorrelation in piecewise regression analysis for mediation analysis SCEDs, we opted to simplify the design. Future research might examine the impact of autocorrelation, missing data, and varying the lengths of the baseline and intervention phases on the performance of MI.

Conclusion

Using mediation analysis to test intervention effects in SCEDs can provide insight into the mechanisms through which interventions achieve their effects for individual participants. This paper evaluated piecewise regression analysis for a single mediator model comparing OLS to three methods for handling autocorrelation, NW, AR(1), and FGLS, and MI under various proportions (20% and 50%) of missing data. The methods were illustrated using data from a walking intervention for osteoarthritis. The simulations indicate that AR(1) and FGLS are promising techniques for modeling autocorrelation, and MI is a promising method for handling missing data in SCEDs for single mediator models. Our results suggest that sample sizes larger than those typically found in SCEDs are recommended to attain acceptable power using the methods evaluated in this study. As the number of tools facilitating data collection continues to rise, larger sample sizes necessary to detect indirect effects in SCEDs using piecewise regression analysis may become more feasible. We hope the results of our simulation studies will contribute to the current scholarship on mediation analysis in SCEDs and promote further research on autocorrelation handling and missing data handling methods in single-case studies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Emma Somer https://orcid.org/0000-0001-9346-3378

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The time index t refers to the measurement occasion, and we assume equidistant time intervals between measurement occasions (e.g., days, weeks). Thus, X_t, M_t, and Y_t are measured at the same occasion (e.g., during the same day). However, we assume that within a measurement occasion (e.g., during a day), the cause variables are measured prior to the effect variables (e.g., X_t is measured before M_t during the same day).
2. RMediation implicitly assumes that the standard errors are consistent, which may not be the case in OLS with nonzero autocorrelation.
3. Note that the proportion of variation accounted for by a_trend, a_level, and the mediator no longer correspond to the intended values of R^2 that led to labeling these effects as large; therefore, the selected effect sizes should no longer be considered as large effects.

References

Albert, J. M., Li, Y., Sun, J., Woyczynski, W. A., & Nelson, S. (2019). Continuous-time causal mediation analysis. *Statistics in Medicine*, *38*(22), 4334–4347. https://doi.org/10.1002/sim.8300

Barlow, D. H., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior for change* (3rd ed.). Pearson/Allyn and Bacon.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, *16*(4), 161–169. https://doi.org/10.1007/s11065-006-9013-7

Bentley, K. H., Kleiman, E. M., Elliott, G., Huffman, J. C., & Nock, M. K. (2019). Real- time monitoring technology in single-case experimental design research: Opportunities and challenges. *Behaviour Research and Therapy*, *117*, 87–96. https://doi.org/10.1016/j.brat.2018.11.017

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist*, *63*(2), 77–95. https://doi.org/10.1037/0003-066X.63.2.77

Borenstein, M, & (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.

Box, G. E. P., & Jenkins, G. M (1970). *Time-series analysis: Forecasting and control*. Holden-Day.

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

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, L. T., Feng, F., Wu, P. J., & Peng, C. J. (2020). Dealing with missing data by EM in single-case studies. *Behavior Research Methods*, *52*(1), 131–150. https://doi.org/10.3758/s13428-019-01210-8

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, *44*(245), 32–61. https://doi.org/10.1080/01621459.1949.10483290

Craig, C. C. (1936). On the frequency function of xy. *Annals of Mathematical Statistics*, *7*(1), 1–15. https://doi.org/10.1214/aoms/1177732541

Davidson, R., & MacKinnon, J. G (2003). *Econometric theory and methods*. Oxford University Press.

Deboeck, P. R., & Preacher, K. J. (2016). No need to be discrete: A method for continuous time mediation analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 61–75. https://doi.org/10.1080/10705511.2014.973960

De, T. K, Michiels, B., Tanious, R., & Onghena, P. (2020). Handing missing data in randomization tests for single-case experiments: A simulation study. *Behavior Research Methods*, *52*(3), 1355–1370. https://doi.org/10.3758/s13428-019-01320-3

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods*, *42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.x

Gaynor, S. T., & Harris, A. (2008). Single-participant assessment of treatment mediators: Strategy description and examples from a behavioral activation intervention for depressed adolescents. *Behavior Modification*, *32*(3), 372–402. https://doi.org/10.1002/cad.20310

Geuke, G., Maric, M., Miočević, M., Wolters, L. H., & de Haan, E. (2019). Testing mediators of youth intervention outcomes using single-case experimental designs. *New Directions for Child and Adolescent Development*, *2019*(167), 39–64. https://doi.org/10.1002/cad.20310

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, *12*(1), 184. https://doi.org/10.1186/1471-2288-12-184

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single- case studies using published studies. *Multivariate Behavioral Research*, *50*(2), 162–183. https://doi.org/10.1080/00273171.2014.973989.

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

Hershberger, M. E., Tuan, R. S., Green, S. B., & Marquis, J. G (1999). Meta-analysis of single-case data. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (Vol. *215*, pp. 107–117). Sage. https://doi.org/10.1006/dbio.1999.9439

Honaker, J., King, G., & Blackwell, M (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. http://dx.doi.org/10.18637/jss.v045.i07

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 79–90. https://doi.org/10.1198/000313007X172556

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, *48*(4), 1294–1304. https://doi.org/10.1093/ije/dyz032

Huitema, B. E., McKean, J. W., & Zhao, J. (1996). The runs test for autocorrelated errors: Unacceptable properties. *Journal of Educational and Behavioral Statistics*, *21*(4), 390–404. https://doi.org/10.3102/10769986021004390

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*(469), 332–346. https://doi.org/10.1198/016214504000001844

Jarrett, M. A., & Ollendick, T. H. (2012). Treatment of comorbid attention-deficit/hyperactivity disorder and anxiety in children: A multiple baseline design analysis. *Journal of Consulting and Clinical Psychology*, *80*(2), 239–244. https://doi.org/10.1037/a0027123

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*(1), 69–86. https://doi.org/10.1207/s15327906mbr2301_4

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

Kleiman, E. M., & Nock, M. K. (2017). Advances in scientific possibilities offered by real-time monitoring technology. *Psychiatry*, *80*(2), 118–124. https://doi.org/10.1080/00332747.2017.1325661

Krasny-Pacini, A., & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine*, *61*(3), 164–179. https://doi.org/10.1016/j.rehab.2017.12.002

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single case designs technical documentation What Works Clearinghouse: Procedures and standards handbook (Version 1.0)*. WWC. http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Lane, J. D., Ledford, J. R., & Gast, D. L. (2017). Single-case experimental design: Current standards and applications in occupational therapy. *American Journal of Occupational Therapy*, *71*(2), Article 7102300010p1–7102300010p9. https://doi.org/10.5014/ajot.2017.022210

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, *171*(5), 624–632. https://doi.org/10.1093/aje/kwp425

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons, Inc.

Lomnicki, Z. A. (1967). On the distribution of products of random variables. *Journal of the Royal Statistical Society*, *29*(3), 513–524. https://doi.org/10.1111/j.2517-6161.1967.tb00713.x

Lutkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer. https://doi.org/10.1007/978-3-540-27752-1

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*(3), 384–389. https://doi.org/10.3758/BF03193007

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104. https://doi.org/10.1037/1082-989x.7.1.83

Mackinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99. https://doi.org/10.1207/s15327906mbr3901_4

Mackinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41. https://doi.org/10.1207/s15327906mbr3001_3

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*(4), 1262–1271. https://doi.org/10.3758/BRM.41.4.1262

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*(2), 201–215. https://doi.org/10.1016/j.jsp.2012.12.005

Maric, M., Wiers, R. W., & Prins, P. J. (2012). Ten ways to improve the use of statistical mediation analysis in the practice of child and adolescent treatment research. *Clinical Child and Family Psychology Review*, *15*(3), 177–191. https://doi.org/10.1007/s10567-012-0114-y

Miočević, M., Klaassen, F., Geuke, G., Moeyaert, M., & Maric, M. (2020). Using Bayesian methods to test mediators of intervention outcomes in single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *14*(1–2), 52–68. https://doi.org/10.1080/17489539.2020.1732029

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 666–683. https://doi.org/10.1080/10705511.2017.1312407

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013a). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*(5), 719–748. https://doi.org/10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013b). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods*, *45*(2), 547–559. https://doi.org/10.3758/s13428-012-0274-1

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. https://doi.org/10.1016/j.jsp.2013.11.003

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal*, *9*(1), 207. https://doi.org/10.32614/RJ-2017-009

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). *Comparison of different methods for univariate time series imputation in r*. arXiv. https://arxiv.org/abs/1510.03924

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), 75. https://doi.org/10.1186/1471-2288-14-75.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

Norell-Clarke, A., Nyander, E., & Jansson-Fröjmark, M. (2011). Sleepless in Sweden: A single subject study of effects of cognitive therapy for insomnia on three adolescents. *Behavioural and Cognitive Psychotherapy*, *39*(3), 367–374. https://doi.org/10.1017/S1352465810000664

O'Brien, N., Philpott-Morgan, S., & Dixon, D. (2016). Using impairment and cognitions to predict walking in osteoarthritis: A series of n-of-1 studies with an individually tailored, data-driven intervention. *British Journal of Health Psychology*, *21*(1), 52–70. https://doi.org/10.1111/bjhp.12153

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., De-Alba, R. G., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large Is large? *School Psychology Review*, *34*(1), 116–132. https://doi.org/10.1080/02796015.2005.12086279

Parker, R. I., & Vannest, K. (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., & Brown, L. (2009). The improvement rate difference for single case research. *Exceptional Children*, *75*(2), 135–150. https://doi.org/10.1177/001440290907500201

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

Peng, C. J., & Chen, L. T. (2018). Handling missing data in single-case studies. *Journal of Modern Applied Statistical Methods*, *17*(1). https://doi.org/10.22237/jmasm/1525133280

Prais, S. J., & Winsten, C. B. (1954, February). *Trend estimators and serial correlation*. (Cowles Commission Discussion Paper No. 383). Retrieved from Cowles Foundation for Research in

Economics website: https://cowles.yale.edu/sites/default/files/files/pub/cdp/s-0383.pdf

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*(1), 99–117. https://doi.org/10.1146/annurev.publhealth.25.102802.124410

Ritter, W. A., Barnard-Brak, L., Richman, D. M., & Grubb, L. M. (2018). The influence of function, topography, and setting on noncontingent reinforcement effect sizes for reduction in problem behavior: A meta-analysis of single-case experimental design data. *Journal of Behavioral Education*, *27*(1), 1–22. https://doi.org/10.1007/s10864-017-9277-4

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989x.7.2.147

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non- overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 163–187. https://doi.org/10.1080/17489530802505412

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

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

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, *37*(6), 1011–1035. https://doi.org/10.1111/cogs.12058

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. https://doi.org/10.1037/a0029312

Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time- series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy*, *43*(3), 679–685. https://doi.org/10.1016/j.beth.2011.10.001

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification*, *34*(3), 195–218. https://doi.org/10.1177/0145445510363306

Springer, M. D., & Thompson, W. E. (1966). The distribution of products of independent random variables. *Siam Journal on Applied Mathematics*, *14*(3), 511–526. https://doi.org/10.1137/0114046

Stefano, S., Quartagno, M., Tamburini, M., & Robinson, D (2018). *orcutt: Estimate procedure in case of first order autocorrelation*. R package version 2.3.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). 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

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., & Wilson, B. (2016). The single-case reporting guideline in BEhavioural interventions (SCRIBE) 2016 statement. *Physical Therapy*, *96*(7), e1–e10. https://doi.org/10.2522/ptj.2016.96.7.e1

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: an R package for mediation analysis confidence intervals. *Behavior Research Methods*, *43*(3), 692–700. https://doi.org/10.3758/s13428-011-0076-x

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24*(5), 637–657. https://doi.org/10.1037/met0000210

van Buuren, S., & Groothuis-Oudshoorn, K (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Velicer, W. F., & Colby, S. M. (2005a). A comparison of missing-data procedures for arima time-series analysis. *Educational and Psychological Measurement*, *65*(4), 596–615. https://doi.org/10.1177/0013164404272502

Velicer, W. F., & Colby, S. M (2005b). Missing data and general transformation approach to time series analysis. In A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 509–535). Erlbaum.

Verbeek, M. (2017). *A guide to modern econometrics*. John Wiley & Sons.

Vilardaga, R., Bricker, J., & McDonell, M. (2014). The promise of mobile technologies and single case designs for the study of individuals in their natural environment. *Journal of Contextual Behavioral Science*, *3*(2), 148–153. https://doi.org/10.1016/j.jcbs.2014.03.003

von Hippel, P. T., & Bartlett, J. W. (2021). Maximum likelihood multiple imputation: Faster imputations and consistent standard errors without posterior draws. *Statistical Science*, *36*(3), 400–420. https://doi.org/10.1214/20-STS793

Wang, L., & Preacher, K. J. (2015). Moderated mediation analysis using Bayesian methods. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 249–263. https://doi.org/10.1080/10705511.2014.935256

Wendt, O. (2009, May). *Calculating effect sizes for single-subject experimental designs: An overview and comparison* [Paper presentation]. The Ninth Annual Campbell Collaboration Colloquium, Oslo, Norway.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. https://doi.org/10.1037/a0016972

Zeileis, A (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1–17. https://doi.org/10.18637/jss.v011.i10

Zeileis, A., & Hothorn, T (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. https://CRAN.R-project.org/doc/Rnews/

Zheng, W., & van der Laan, M. (2017). Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of Causal Inference*, *5*(2), Article 20160006. https://doi.org/10.1515/jci-2016-0006

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes i: Building a general cross-lagged panel model (gclm). *Organizational Research Methods*, *23*(4), 651–687. https://doi.org/10.1177/1094428119847278