Dealing with missing data by EM in single-case studies

Li-Ting Chen¹ · Yanan Feng² · Po-Ju Wu² · Chao-Ying Joanne Peng³

Published online: 25 February 2019
© The Psychonomic Society, Inc. 2019

Abstract

Single-case experimental design (SCED) research plays an important role in establishing and confirming evidence-based practices. Due to multiple measures of a target behavior in such studies, missing information is common in their data. The expectation–maximization (EM) algorithm has been successfully applied to deal with missing data in between-subjects designs, but only in a handful of SCED studies. The present study extends the findings from Smith, Borckardt, and Nash (2012) and Velicer and Colby (2005b, Study 2) by systematically examining the performance of EM in a baseline–intervention (or AB) design under various missing rates, autocorrelations, intervention phase lengths, and magnitudes of effects, as well as two fitted models. Three indicators of an intervention effect (baseline slope, level shift, and slope change) were estimated. The estimates' relative bias, root-mean squared error, and relative bias of the estimated standard error were used to assess EM's performance. The findings revealed that autocorrelation impacted the estimates' qualities most profoundly. Autocorrelation interacted with missing rate in impacting the relative bias of the estimates, impacted the root-mean squared error nonlinearly, and interacted with the fitted model in impacting the relative bias of the estimated standard errors. A simpler model without autocorrelation can be used to estimate baseline slope and slope change in time-series data. EM is recommended as a principled method to handle missing data in SCED studies. Two decision trees are presented to assist researchers and practitioners in applying EM. Emerging research directions are identified for treating missing data in SCED studies.

Keywords Missing data · Single-case designs · Autocorrelation · Phase and slope change · EM · Effect size

SCED studies play an important role in establishing and confirming evidence-based interventions (Horner et al., 2005; Shadish & Sullivan, 2011; Smith, 2012). SCEDs have been used in a variety of disciplines, including education, counseling, psychology, psychiatry, neuroscience, medicine, social work, and sport sciences (Gustafson, Nassar, & Waddell, 2011; Kratochwill, 2015; Richards, Taylor, & Ramasamy, 2014). SCED studies are characterized by repeated measurements of a target behavior from a small number of participants. Although there are variations of SCEDs, SCED studies in general consist of a baseline phase and an intervention phase in which several observations of the target behavior take place. SCED data are usually correlated in a time-forward

fashion. For example, an observation at time t may be correlated with its subsequent observations. If a behavior, and therefore its measurement, stabilizes during the baseline phase, it is said to be interrupted when an intervention is introduced in the intervention phase. This pattern of data can be modeled by a lag-1 interrupted time-series model, where *lag-1* refers to adjacent data and the time-series model allows for adjacent data to be correlated. Due to repeated measurements of the same behavior(s) in a participant over a period of time, missing portions of data are a common occurrence in SCED studies (Allison, Silverstein, & Gorman, 1996). This type of missing data is referred to as *item-level missing* in the literature (Dong & Peng, 2013; Newman, 2014).

To assess the occurrence of missing data in current SCED studies, we reviewed articles published in 2015 and 2016 in five peer-reviewed journals (*Behavior Modification*, *Journal of Applied Behavior Analysis*, *Journal of Positive Behavior Interventions*, *Journal of School Psychology*, and *Journal of Special Education*). These five journals aim to publish behavioral analysis studies in clinical or school settings. Approximately 18% (or 33 out of 182) of the SCED articles had missing data. Another 9% (or 17 out of 182) had

✉ Li-Ting Chen
litingc@unr.edu

¹ University of Nevada, Reno, NV, USA

² Indiana University, Bloomington, IN, USA

³ National Taiwan University, Taipei, Taiwan

insufficient information to determine whether missing data existed. The lowest missing-data rate was 1%, and the highest was 45%. In general, there was a higher missing rate in the intervention phase than in the baseline phase. The average missing rate for the baseline data was 6%, and the average missing rate for the intervention data was 15%. Because serious consequences can result from inadequate treatment of missing data, even by specialized computing tools (Chen, Peng, & Chen, 2015; Velicer & Colby, 2005a), these findings speak to the need to critically evaluate the application of missing data methods in SCED studies.

According to Chen et al. (2015), computing tools specialized in analyzing SCED data have handled missing data in one of three ways: (a) disallowing missing data from any participant, (b) using available data only, or (c) replacing missing scores with 0. If the *a* approach is adopted, the analysis is terminated when it encounters missing scores, or such participants' data are flagged as errors (Chen et al., 2015). If the *b* approach is applied and missing data occurred during the last three sessions of a baseline phase or the first three sessions of an intervention phase, researchers would be unable to fully examine the immediacy of the intervention effect (What Works Clearinghouse, 2017). Furthermore, removing missing data section(s) inevitably alters the study design and creates difficulty in integrating the results across participants or studies. The literature has shown that the *b* approach leads to overestimating the slope in time-series data (Velicer & Colby, 2005a). The *c* approach is unacceptable, because a missing score is not necessarily equal to zero. Replacing missing scores with the mean of the available scores in the same phase is one popular ad hoc method, called *mean substitution*. However, mean substitution produces biased parameter estimates in between-subjects designs (Little & Rubin, 1987, 2002; Peng, Harwell, Liou, & Ehman, 2006), as well as in time-series designs such as SCEDs (Velicer & Colby, 2005a; Velicer & Fava, 2003).

Using a simulation approach, Velicer and Colby (2005b, Study 2) investigated the performance of the maximum likelihood method under two missing-data conditions: missing at random (MAR) and missing completely at random (MCAR). The MAR condition assumes that the probability of missing data may be a function of the data observed, but not a function of the data missing (Little & Rubin, 1987). MAR is less stringent and more realistic than MCAR (Little & Rubin, 1987). MCAR assumes that the probability of missing is a function of neither the data observed nor the data missing. Hence, the incomplete data under MCAR represent a random sample of the population, if they are randomly selected

Table 1. ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RB for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a short intervention phase ($n_A = 10$, $n_B = 10$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.00108565	3,576.68	< .0001	.47*
Autocorrelation (A)	18	0.00008447	278.29	< .0001	.33*
Model fitting (MF)	1	0.00000198	6.53	.0150	< .06
MR*A	36	0.00002532	83.42	< .0001	.20*
A*MF	18	0.00000067	2.20	.0216	< .06
MR*MF	2	0.00000628	20.70	< .0001	< .06
Error	36	0.00000030			
β_2 (level shift)					
Missing rate (MR)	2	0.04192584	716.73	< .0001	.39*
Autocorrelation (A)	18	0.00255364	43.65	< .0001	.22*
Model fitting (MF)	1	0.00011456	1.96	.1702	< .06
MR*A	36	0.00217570	37.19	< .0001	.37*
A*MF	18	0.00002136	0.37	.9872	< .06
MR*MF	2	0.00120111	20.53	< .0001	< .06
Error	36	0.00005850			
β_3 (slope change)					
Missing rate (MR)	2	0.00185399	3,924.56	< .0001	.45*
Autocorrelation (A)	18	0.00015087	319.36	< .0001	.33*
Model fitting (MF)	1	0.00000441	9.34	.0042	< .06
MR*A	36	0.00004753	100.61	< .0001	.21*
A*MF	18	0.00000125	2.64	.0064	< .06
MR*MF	2	0.00000218	4.62	.0164	< .06
Error	36	0.00000047			

Boldface β_2 (level shift) denotes that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

(Little & Rubin, 1987). Velicer and Colby (2005b, Study 2) simulated data from a lag-1 interrupted time-series model with 50 baseline scores and 50 intervention scores. The MCAR condition was created by randomly eliminating a certain proportion of scores. The MAR condition was created by removing every other score, or consecutive scores (Velicer & Colby, 2005b, Study 2). For example, in the 40% missing condition, every-other deletion deleted the odd-numbered scores from the 11th to the 50th in the baseline (or A) phase, and from the 61st to the 100th in the intervention (or B) phase. In the same 40% missing condition, consecutive deletion deleted the scores from the 16th to the 35th in the A phase and from the 66th to 85th in the B phase. A total of four variables were manipulated: missing-data pattern (with three levels: MCAR, MAR created by removing every other score, or MAR created by removing consecutive scores), missing-data proportion (with three levels: 0%, 20%, or 40%),

Fig. 1 RBs of β_2 (level shift) when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 10$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RBs are those below the bold reference line at $RB = -0.05$ or above the bold reference line at $RB = 0.05$

model specification (with two levels: correct or incorrect time-series model), and autocorrelation (with two levels: .6 or -.6).

The results from Velicer and Colby (2005b, Study 2) revealed that correct estimates for the parameters were obtained in all manipulated conditions. However, these results were obtained from a handful of replications (= 10) over 100 scores. The length of the A and B phases was large, as compared with most published SCED studies (Shadish & Sullivan, 2011). Furthermore, the missing proportions were kept the same for both the A and B phases under MAR. Yet our review uncovered differential missing patterns—that is, the B phase tended to have a higher missing rate (15% on average) than the A phase (6% on average). Finally, time-series modeling requires a model specification that may be challenging to practitioners not trained in time-series modeling techniques.

In recent years, expectation–maximization (EM) has been suggested as a viable option for dealing with missing data in SCED (e.g., Allison et al., 1996; Smith, Borckardt,

& Nash, 2012). EM was proposed by Dempster, Laird, and Rubin (1977) to handle missing data under MAR based on maximum-likelihood estimation of parameters, such as means and variances. The EM method is an iterative process (Little & Rubin, 2002). Each iteration of EM consists of an expectation (E) step and a maximization (M) step. In the E step, the conditional expectation of the complete data log-likelihood is derived in the presence of the observed data and the current estimates for parameters. In the M step, the conditional expectation of the complete data log-likelihood is maximized in order to yield a new set of parameter estimates. The E and M steps iterate until the difference in the observed log-likelihood from two consecutive iterations meets a prespecified convergence criterion, such as $1E-4$ in SAS, version 14.1 (SAS Institute Inc., 2015). When the EM algorithm converges according to its convergence criterion, a final set of estimates for parameters (e.g., means, variances/covariances) are obtained. From these estimates, expected values for the missing data can be derived from the EM

Table 2 ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RMSE for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a short intervention phase ($n_A = 10$, $n_B = 10$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.00363232	1,239.12	< .0001	.06*
Autocorrelation (A)	18	0.00652217	2,224.94	< .0001	.92*
Model fitting (MF)	1	0.00018147	61.90	< .0001	< .06
MR*A	36	0.00005949	20.29	< .0001	< .06
A*MF	18	0.00002974	10.14	< .0001	< .06
MR*MF	2	0.00002178	7.43	.0020	< .06
Error	36	0.00000293			
β_2 (level shift)					
Missing rate (MR)	2	0.27604465	1,442.90	< .0001	.11*
Autocorrelation (A)	18	0.23354083	1,220.73	< .0001	.86*
Model fitting (MF)	1	0.02554601	133.53	< .0001	< .06
MR*A	36	0.00198553	10.38	< .0001	< .06
A*MF	18	0.00277300	14.49	< .0001	< .06
MR*MF	2	0.00290281	15.17	< .0001	< .06
Error	36	0.00019131			
β_3 (slope change)					
Missing rate (MR)	2	0.00720776	14,163.5	< .0001	< .06
Autocorrelation (A)	18	0.01713576	33,672.4	< .0001	.94*
Model fitting (MF)	1	0.00003584	70.43	< .0001	< .06
MR*A	36	0.00014149	278.03	< .0001	< .06
A*MF	18	0.00000400	7.87	< .0001	< .06
MR*MF	2	0.00000577	11.35	.0002	< .06
Error	36	0.00000051			

Boldface β_2 (level shift) denotes that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

algorithm. EM is easy to implement and stable (Couvreur, 1997), and is a principled method that has been shown to yield unbiased estimates of parameters when its assumptions (i.e., multivariate normality and MAR) are met (Barnard, 2000; Schafer, 1997). Furthermore, EM is more efficient, in terms of smaller standard error of parameter estimates, than the multiple-imputation method, because EM is based on the observed data log-likelihood and does not require simulations of multiple data sets (Dong & Peng, 2013; Springer & Urban, 2014). EM is also available from popular statistical software, such as SPSS, SAS, Stata, and R.

Smith et al. (2012) evaluated the performance of EM under various proportions of missing data and autocorrelations on data mimicked from “a typical 8-week study of psychotherapy effectiveness” (p. 681). Data were simulated for an AB design in which the length of the A phase

was 10 and that of the B phase was 56, according to an interrupted time-series model with lag-1 autocorrelation specified at one of five levels: .0, .2, .4, .6, or .8. For each of the 5,000 replications, data were deleted under MCAR in one of four proportions: 10%, 20%, 30%, or 40%. The missing data were subsequently replaced using the EM algorithm implemented in PASW Statistics for Windows and Mac (SPSS Inc., 2010; Version 18.0.0). The replaced and the nonmissing data together formed completed data that were subsequently analyzed to determine power sensitivity under each joint condition of missing proportion and autocorrelation. Power sensitivity in Smith et al. was defined as the fraction of the number of significant level-change effects correctly identified over the number of significant level-change effects correctly and incorrectly identified. The findings from Smith et al. showed that the power sensitivity was not affected by missing data proportions, even when the proportion was 40%. Furthermore, when the autocorrelation was .8, power sensitivity was lower than the acceptable level of .80, regardless of the missing proportion. Thus, Smith et al. provided support for using EM to accurately estimate the level-change effect in the presence of missing data.

However, six aspects of Smith et al.’s (2012) study render their findings limited in generalizability and interpretation. First, their missing data were created under MCAR for each replication. As we previously noted, under MCAR, the nonmissing data were a random representation of their underlying population, and therefore could be used to estimate the unknown population parameters (Schafer & Graham, 2002; Schlomer, Bauman, & Card, 2010). It is therefore not surprising that Smith et al. (2012) did not find the missing proportion to have impacted power sensitivity. Second, the missing proportion (e.g., 40%) in Smith et al. was implemented for both the A and B phases. It was possible in their simulation study to incur missing scores either only in one phase or evenly in both. These scenarios could not have adequately reflected the real-world missing-data patterns uncovered in our review of SCED studies, previously discussed. Third, although a short A phase was specified by Smith et al., a short B phase was not investigated. According to Shadish and Sullivan (2011), a B phase shorter than 56 was quite common. Their review of 113 SCED studies published in 2008 revealed that the median data points per case were 20. Ferron, Farmer, and Owens (2010) reported that the data points per case ranged from seven to 58, with a median of 24. Ferron et al. also reported that the A phase length ranged from three to ten data points, with a median of four. Hence, a shorter B phase with, say, ten scores, has yet to be examined. Fourth, only positive autocorrelations were investigated, yet negative autocorrelations have been reported in empirical SCED studies

Fig. 2 RMSEs of β_2 (level shift), when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 10$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RMSEs are those above the bold reference line at $RMSE = 1$

(Harrington & Velicer, 2015; Parker et al., 2005; Shadish & Sullivan, 2011). For example, Shadish and Sullivan's review uncovered lag-1 autocorrelations ranging from $-.93$ to $.79$. Additionally, studies examining the performance of missing-data methods with time-series data have shown different results under positive and negative autocorrelations (Velicer & Colby, 2005a, 2005b). Fifth, the main effects were interpreted graphically, and the interaction between missing-data proportion and autocorrelation was not examined. Sixth, only one measure of impact—namely, power sensitivity—was analyzed and interpreted. Due to these limitations, the findings from Smith et al.'s study are not readily generalizable to other SCED studies with different intervention phase lengths and/or data characteristics.

The present simulation study extended and improved on Smith et al. (2012) and Velicer and Colby (2005b, Study 2) in the following respects, in order to systematically investigate the performance of EM as a missing-data method for SCED studies.

1. Missing data were simulated under the MAR condition.
2. Higher missing-data rates (i.e., 10%, 20%, and 30%) were specified for the B phase than for the A phase (i.e., 0%, 10%, and 20%), to mimic differential missing rates uncovered in empirical studies for the two phases.
3. A short B phase with ten scores was included, along with a long B phase with 56 scores, as in Smith et al. (2012) and similarly in Velicer and Colby (2005b, Study 2). The A phase was fixed at ten scores, as in Smith et al.'s study. The combination of a long A phase and a short B phase was not included, because such a combination was not found empirically in Shadish and Sullivan's (2011) or Ferron et al.'s (2010) review. An A phase shorter than ten, but with at least five data points, as recommended by the What Works Clearinghouse (WWC) Standards (WWC, 2017), was not considered either, because such a short A phase would not allow us to implement exactly 10% and 20% missing rates for the A phase.

Table 3 ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RBESE for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a short intervention phase ($n_A = 10$, $n_B = 10$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.19962535	26.83	< .0001	< .06
Autocorrelation (A)	18	0.91912125	123.54	< .0001	.60*
Model fitting (MF)	1	3.05994960	411.29	< .0001	.11*
MR*A	36	0.00048634	0.07	1.0000	< .06
A*MF	18	0.39844053	53.56	< .0001	.26*
MR*MF	2	0.03150585	4.23	.0223	< .06
Error	36	0.00743984			
β_2 (level shift)					
Missing rate (MR)	2	0.19135029	27.27	< .0001	< .06
Autocorrelation (A)	18	0.85707479	122.15	< .0001	.60*
Model fitting (MF)	1	2.93198046	417.86	< .0001	.11*
MR*A	36	0.01774337	< .01	.07	< .06
A*MF	18	0.36559613	52.10	< .0001	.26*
MR*MF	2	0.03472999	4.95	.0126	< .06
Error	36	0.00701661			
β_3 (slope change)					
Missing rate (MR)	2	0.20766479	25.60	< .0001	< .06
Autocorrelation (A)	18	1.02659538	126.55	< .0001	.60*
Model fitting (MF)	1	3.48231959	429.28	< .0001	.11*
MR*A	36	0.00046060	0.06	1.0000	< .06
A*MF	18	0.45952703	56.65	< .0001	.27*
MR*MF	2	0.03150907	3.88	.0297	< .06
Error	36	0.00811207			

Boldface β_1 (baseline slope), β_2 (level shift), and β_3 (slope change) denote that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

- Both negative and positive autocorrelations were investigated.
- A simplified model without autocorrelation was fitted to the data, along with a lag-1 interrupted time-series model that was used to simulate the data.
- Three effects—namely, slope of the baseline phase, level shift, and slope change of the time-series model—were varied from small to large.
- The number of replications under each condition was increased to 10,000, from ten in Velicer and Colby (2005b, Study 2) or 5,000 in Smith et al. (2012), to ensure the stability of the results.
- Multiple criteria were used to assess the singular and joint impacts of missing-data rate, autocorrelation, intervention phase length, magnitude of the effects, and model fitting on the performance of EM.

In sum, in the present study we aimed to answer the research question of how missing rate, the magnitude and direction of autocorrelation, intervention phase length, the magnitude of effects, and model fitting impacted the performance of EM as a missing-data method. The AB design adopted for this simulation study is fundamental, or has served as a building block, to all other designs used in intervention studies. The findings from the present study are applicable and generalizable to other SCED studies consisting of multiple baseline and intervention phases (e.g., ABAB designs), or to intervention phases starting at staggered time points for different participants (e.g., multiple-baseline designs; WWC, 2017). The results should also provide empirical evidence to support the application of a simpler model to fitted time-series data, without estimating autocorrelation. It is our hope that the present study will prompt additional research on and wider application of advanced missing-data methods for the analysis of SCED data.

Method

Outline of simulation and analysis procedures

The simulation and analysis procedures are outlined in five steps below. Additional details on each step are presented following this outline.

Step 1: A total of 10,000 replications were simulated from a time-series model with prespecified autocorrelation and model parameters to mimic data from an AB design. Twenty scores were simulated for the A phase ($n_A = 10$) and a short B phase ($n_B = 10$), whereas 66 scores were simulated for the A phase ($n_A = 10$) and a long B phase ($n_B = 56$).

Step 2: Missing data were created under MAR at a low, medium, or high missing rate: (1) low missing rate = 0% in the A phase and 10% in the B phase; (2) medium missing rate = 10% in the A phase and 20% in the B phase; and (3) high missing rate = 20% in the A phase and 30% in the B phase.

Step 3: The EM algorithm from the SAS system for Windows, version 14.1, was employed to treat the missing data created in Step 2.

Step 4: The completed data were subsequently fitted by a correct time-series model that was used to simulate the data in Step 1, and also by a simplified model without autocorrelation.

Step 5: The impacts of missing rate, autocorrelation, and model fitting were assessed separately for short and long B phases and for different magnitudes of the model parameters, based on the three criteria.

Fig. 3 RBESes of β_2 (level shift), when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 10$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RBESes are those below the bold reference line at RBESE = -0.1 or those above the bold reference line at RBESE = 0.1

The entire SAS program is available in the supplemental materials, at <https://nevada.box.com/shared/static/7aplqc9d5tbx1dadglouu3q1rsrk232i.docx>. The SAS program was written by the second author. The accuracy of the SAS program was verified by the first and the third authors.

Step 1: Simulating data from a time-series model

The time-series model suggested by Huitema and McKean (2000), Manolov and Solanas (2008), and Manolov, Solanas, Sierra, and Evans (2011) was used to simulate each replication. The model is expressed in Eq. 1 with three parameters (β_1 to β_3) known to be indicators of an intervention effect (Allison & Gorman, 1993; Wong, Wing, Steiner, Wong, & Cook, 2012):

$$Y_t = \beta_1 \times T_t + \beta_2 \times D_t + \beta_3 \times SC_t + \varepsilon_t, \quad (1)$$

where Y_t = the observed score at time t ; β_1 = baseline slope = 0.1 or 1 ; $T_t = t = 1, \dots, N$ ($n_A + n_B = 20$ or 66); β_2 = level shift = 0.1 or 1 (= the difference between two expected scores at the $n_A + 1$ time point, one expected score was derived from the baseline slope, and the other from the intervention slope); $D_t = 0$ for $t = 1, \dots, n_A$ (the baseline scores), and $D_t = 1$ for $t = n_A + 1, \dots, N$ (the intervention scores); β_3 = slope change from the baseline to the intervention slope = 0.1 or 1 ; $SC_t = 0$ for $t = 1, \dots, n_A$ (the baseline scores), and $SC_t = 0, 1, \dots, n_B - 1$ for $t = n_A + 1, \dots, N$ (the intervention scores); and ε_t = error term.

Guided by Manolov and Solanas (2008) and Smith et al. (2012), we generated $\varepsilon_t = \phi_1 \times \varepsilon_{t-1} + u_t$, where ϕ_1 is the autocorrelation, and u_t is noise randomly generated from a standard normal distribution with a mean of 0 and an SD of 1 . At $t = 1$, ε_1 was set to equal u_1 . The variance of error terms (ε_t) was standardized to 1 at each time point (Huitema & McKean, 2007; Solanas, Manolov, & Onghena, 2010).

Table 4 ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RB for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a long intervention phase ($n_A = 10$, $n_B = 56$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.00600505	5,151.53	< .0001	.71*
Autocorrelation (A)	18	0.00014550	124.82	< .0001	.16*
Model fitting (MF)	1	0.00000055	0.47	.4964	< .06
MR*A	36	0.00005891	50.54	< .0001	.13*
A*MF	18	0.00000320	2.74	.0049	< .06
MR*MF	2	0.00001550	13.30	< .0001	< .06
Error	36	0.00000117			
β_2 (level shift)					
Missing rate (MR)	2	0.40791575	8,625.29	< .0001	.77*
Autocorrelation (A)	18	0.00851640	180.08	< .0001	.15*
Model fitting (MF)	1	0.00001399	0.30	.5898	< .06
MR*A	36	0.00218869	46.28	< .0001	.07*
A*MF	18	0.00017367	3.67	.0004	< .06
MR*MF	2	0.00051134	10.81	.0002	< .06
Error	36	0.00004729			
β_3 (slope change)					
Missing rate (MR)	2	0.00628882	5,339.79	< .0001	.73*
Autocorrelation (A)	18	0.00011997	101.86	< .0001	.13*
Model fitting (MF)	1	0.00000050	0.43	.5169	< .06
MR*A	36	0.00006440	54.68	< .0001	.13*
A*MF	18	0.00000320	2.72	.0052	< .06
MR*MF	2	0.00001466	12.45	< .0001	< .06
Error	36	0.00000118			

Boldface β_2 (level shift) denotes that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

Three model parameters (i.e., β_1 , β_2 , and β_3) were varied from .1 (small effect = 10% of the *SD* of scores at each time point) to 1 (large effect = 100% of the *SD* of scores at each time point). Four magnitudes of the β s were considered in the present study: (1) $\beta_1 = \beta_2 = \beta_3 = .1$ (all effects were small), (2) $\beta_1 = 1$, $\beta_2 = \beta_3 = .1$ (baseline slope was large while the other two effects were small), (3) $\beta_1 = \beta_3 = .1$, $\beta_2 = 1$ (level shift was large while the other two effects were small), and (4) $\beta_1 = \beta_2 = .1$, $\beta_3 = 1$ (slope change was large while the other two effects were small).

Step 2: Creating missing data under MAR

Given a specific missing rate (e.g., 0% in A and 10% in B) and a specific time t , the score (Y_t) was made missing under MAR, depending on the dummy-coded D_t in Eq. 1. The missing rate was then implemented through a

binomial parameter. When the missing rate was 0% in A and 10% in B, the parameter of the binomial random variable was 0 for all $D_t = 0$ (A phase) scores, and .1 for all $D_t = 1$ (B phase) scores. Given the specific parameter, the binomial random variable took on a value of 1 or 0. When the binomial random variable = 1, the corresponding Y_t became missing. When the binomial random variable = 0, its corresponding Y_t remained unchanged or nonmissing. Thus, the probability of a score being missing was not conditioned on the score itself, but on the dummy-coded D_t . At the conclusion of this step, we verified that the rates of the missing scores matched the low, medium, and high missing rates prespecified.

Step 3: Treating missing data by EM

The EM algorithm in PROC MI of SAS Version 14.1 was employed to treat missing scores at each time point. The option MAXITER = 500 was specified, to set the maximum number of iterations in EM to 500. After the EM algorithm converged, final estimates of the means, variances, and covariances for all variables (i.e., Y , T , D , SC) in Eq. 1 were obtained. From these estimates, expected values for the missing Y scores were derived from a linear combination of T , D , and SC . The expected values replaced the missing Y scores. The replaced Y scores and the nonmissing scores formed a completed data set, stored in an output data set for subsequent analyses (SAS Institute Inc., 2015).

Step 4: Model fitting with and without autocorrelation

The completed data were subsequently fitted by a time-series model with autocorrelation (Eq. 1) and by a simplified model without autocorrelation, in which the errors were assumed to be independent. The β s in Eq. 1 for the time-series model were estimated using PROC AUTOREG in SAS Version 14.1, and the β s for the simplified model were estimated using PROC IML in SAS Version 14.1. These estimates were evaluated using the three criteria.

Step 5: Assessing the impact by three criteria

Three criteria were used to assess the impacts of missing rate, autocorrelation, intervention phase length, the magnitude of effects, and model fitting on estimates of the three β s: relative bias (RB), root-mean squared error (RMSE), and relative bias of the estimated standard error (RBSE). For these three criteria, the smaller the absolute values, the better were the corresponding estimates. RB assessed the accuracy of the estimates,

Fig. 4 RBs of β_2 (level shift), when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 56$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RBs are those below the bold reference line at $RB = -0.05$ or above the bold reference line at $RB = 0.05$

relative to the true values. RMSE assessed the precision of the estimates; it measured both bias and sampling variance, when the estimate was biased (Enders, 2010). And RBESE assessed the bias in SE . If SE were biased downward or underestimated on average, its corresponding test statistic (i.e., t) could be overestimated. An overestimated test statistic leads to a liberal test that rejects a correct H_0 of no intervention effect more probably than it should. If SE were biased upward or overestimated on average, its corresponding test statistic could be underestimated. An underestimated test statistic leads to a conservative test that rejects an incorrect H_0 less probably than it should, due to reduced statistical power. Because the population SE could not be derived analytically in the present study, we used the SE obtained from fitting the correct time-series model to the complete data generated in Step 1 as a proxy for the population SE (Hoogland & Boomsma, 1998).

Let θ be the population effect or the parameter of interest, whether it was β_1 , β_2 , or β_3 . Let $\hat{\theta}_i$ be an estimate of θ , computed from each of the 10,000 ($= R$) replications. RB, RMSE, and RBESE are defined in Eqs. 2, 3, and 4, respectively:

$$RB = \frac{\bar{\hat{\theta}}_i - \theta}{\theta}, \quad (2)$$

where $\bar{\hat{\theta}}_i$ is the mean of the R estimates for θ_i under each condition;

$$RMSE = \sqrt{\frac{\sum_{i=1}^R (\hat{\theta}_i - \theta)^2}{R}}; \text{ and} \quad (3)$$

$$RBESE = \frac{\overline{\widehat{SE}_i} - \overline{\widehat{SE}_c}}{\widehat{SE}_c}, \quad (4)$$

Table 5 ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RMSE for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a long intervention phase ($n_A = 10$, $n_B = 56$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.00361244	1,235.36	< .0001	.06*
Autocorrelation (A)	18	0.00604652	2,067.74	< .0001	.91*
Model fitting (MF)	1	0.00084690	289.62	< .0001	< .06
MR*A	36	0.00004230	14.47	< .0001	< .06
A*MF	18	0.00004027	13.77	< .0001	< .06
MR*MF	2	0.00005047	17.26	< .0001	< .06
Error	36	0.00000292			
β_2 (level shift)					
Missing rate (MR)	2	0.17305539	304.90	< .0001	.06*
Autocorrelation (A)	18	0.29160485	513.77	< .0001	.87*
Model fitting (MF)	1	0.13008746	229.20	< .0001	< .06
MR*A	36	0.00120002	2.11	.0137	< .06
A*MF	18	0.01175591	20.71	< .0001	< .06
MR*MF	2	0.00667712	11.76	.0001	< .06
Error	36	0.00056758			
β_3 (slope change)					
Missing rate (MR)	2	0.00357725	1,333.42	< .0001	.06*
Autocorrelation (A)	18	0.00658344	2,453.97	< .0001	.92*
Model fitting (MF)	1	0.00071530	266.63	< .0001	< .06
MR*A	36	0.00004415	16.46	< .0001	< .06
A*MF	18	0.00003456	12.88	< .0001	< .06
MR*MF	2	0.00004452	16.60	< .0001	< .06
Error	36	0.00000268			

Boldface β_2 (level shift) denotes that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

where $\overline{SE}_i$ was the average of 10,000 estimated standard errors (ESEs) obtained from Step 4 for a given manipulated condition, and $\overline{SE}_c$ was the average of 10,000 ESEs obtained from fitting the correct time-series model to complete data generated from Step 1.

Scope of this simulation study

Altogether, 912 conditions were simulated and analyzed. The 912 conditions resulted from manipulating (1) three missing rates, (2) 19 autocorrelations ($\phi_1 = -.9$ to $+.9$ in increments of .1), (3) two intervention phase lengths (10 for a short B phase, and 56 for a long B phase), (4) four magnitudes of the β s ($\beta_1 = \beta_2 = \beta_3 = 0.1$; $\beta_1 = 1$, $\beta_2 = \beta_3 = 0.1$; $\beta_1 = \beta_3 = 0.1$, $\beta_2 = 1$; and $\beta_1 = \beta_2 = 0.1$, $\beta_3 = 1$), and (5) two models fitted to the data (a time-series model and a simplified model). Additionally, the complete data (i.e., 0% missing in both the A and B phases) were simulated for 152 combinations of the 19

autocorrelations, two intervention phase lengths, and four magnitudes of the β s. Each complete data set was fitted by the correct time-series model (Eq. 1) in order to obtain a proxy for the population SE (Hoogland & Boomsma, 1998).

Results

Three-way analyses of variance (ANOVAs) were carried out to examine the impacts of missing rate, autocorrelation, and model fitting on the RB, RMSE, and RBESE for each combination of two intervention phase lengths and four magnitudes of the three β s. Thus, 72 [= 3 (criteria) $\times$ 2 (intervention phase lengths) $\times$ 4 (magnitudes of β s) $\times$ 3 (β s)] three-way ANOVAs were performed. For each three-way ANOVA, only one RB (or RMSE or RBESE) was computed from the 10,000 replications. Consequently, the three-way interaction was used as an error term to statistically test the three main effects of missing rate, autocorrelation, and model fitting and their two-way interactions. A statistical test result was deemed significant, in the present study, if its significance level (p) was less than .05 and its η^2 , as an effect size measure, was at least .06—a medium effect size, according to Cohen (1969). Otherwise, a statistical test result was deemed nonsignificant.

Furthermore, acceptable RBs, RMSEs, and RBESEs were identified in order to fully understand the performance of EM under the manipulated conditions. We applied Hoogland and Boomsma's (1998) cutoffs for acceptable RB and RBESE. According to Hoogland and Boomsma, $|RB| < .05$ is acceptable and $|RB| \geq .05$ is unacceptable, and $|RBESE| < .1$ is acceptable and $|RBESE| \geq .1$ is unacceptable. As for RMSE, we defined $RMSE < 1$ as being acceptable and $RMSE \geq 1$ as being unacceptable. The cutoff of 1 for acceptable RMSE was established in order to identify those RMSEs that were no larger than the standardized error variance at each time point.

The results obtained from $\beta_1 = \beta_2 = \beta_3 = 0.1$, paired with a short intervention phase ($n_A = 10$, $n_B = 10$) are presented first, in the order of RB, RMSE, and RBESE. The results obtained from $\beta_1 = \beta_2 = \beta_3 = 0.1$ paired with a long intervention phase ($n_A = 10$, $n_B = 56$) are presented next, followed by the results obtained from other magnitudes of the β s. Conclusions based on all results are presented last. The results not shown or discussed in this article are available in the supplemental materials.

$\beta_1 = \beta_2 = \beta_3 = 0.1$ paired with a short intervention phase ($n_A = 10$, $n_B = 10$)

RB Table 1 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RB for β_1 , β_2 , and β_3 . The main effects of

Fig. 5 RMSEs of β_2 (level shift), when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 56$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RMSEs are those above the bold reference line at RMSE = 1

missing rate and autocorrelation and their interaction yielded significant results for the three β s. The model fitting did not significantly affect the RBs of the three β s. The RBs of the β_1 and β_3 estimates were all acceptable; their graphs are presented in the supplemental materials. Some of the RBs for β_2 were unacceptable, depending on the missing rate and the magnitude of autocorrelation. Boldface font in Table 1 identifies significant results combined with unacceptable RBs.

Figure 1 displays the RBs of β_2 obtained from all conditions manipulated. Specifically, the circles, triangles, and squares, respectively, represent the RBs obtained from three missing rates: 0% in the A phase and 10% in the B phase (circles), 10% in A and 20% in B (triangles), and 20% in A and 30% in B (squares). The solid circles, triangles, and squares represent the RBs obtained from fitting the time-series model, and the hollow shapes represent the RBs obtained from fitting the simplified model.

According to Fig. 1, the patterns of RBs for β_2 are not parallel across different missing rates, due to the significant interaction of missing rate with autocorrelation. The RBs obtained from the two models did not show noticeable differences, confirming the nonsignificant effect of the model fitting. Specifically, when the missing rate = 0% in A and 10% in B, the RBs were unacceptable from $\phi_1 = .5$ to $.9$. When the missing rate = 10% in A and 20% in B, the RBs were unacceptable from $\phi_1 = -.2$ to $.6$. When the missing rate = 20% in A and 30% in B, the RBs were unacceptable when $\phi_1 = -.9$ to $-.8$.

RMSE Table 2 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RMSE for β_1 , β_2 , and β_3 . The main effects of missing rate and autocorrelation yielded significant results for β_1 and β_2 . For β_3 , only the main effect of autocorrelation was significant. None of the two-way interactions was significant. The large η^2 s ($\geq .86$) due to

Table 6 ANOVA results of missing rate, autocorrelation, model fitting, and their interactions on RBESE for β_1 , β_2 , and β_3 when $\beta_1 = \beta_2 = \beta_3 = .1$ paired with a long intervention phase ($n_A = 10$, $n_B = 56$)

Source	df	MS	F	p	η^2
β_1 (baseline slope)					
Missing rate (MR)	2	0.13059396	17.53	< .0001	< .06
Autocorrelation (A)	18	1.09968888	147.64	< .0001	.65*
Model fitting (MF)	1	1.75048581	235.01	< .0001	.06*
MR*A	36	0.00037386	0.05	1.0000	< .06
A*MF	18	0.45636685	61.27	< .0001	.27*
MR*MF	2	0.03006158	4.04	.0262	< .06
Error	36	0.00744854			
β_2 (level shift)					
Missing rate (MR)	2	0.12339806	16.58	< .0001	< .06
Autocorrelation (A)	18	1.11285552	149.54	< .0001	.64*
Model fitting (MF)	1	2.04679561	275.03	< .0001	.07*
MR*A	36	0.00037752	0.05	1.0000	< .06
A*MF	18	0.48384552	65.01	< .0001	.28*
MR*MF	2	0.04019848	5.40	.0089	< .06
Error	36	0.00744209			
β_3 (slope change)					
Missing rate (MR)	2	0.13220004	17.41	< .0001	< .06
Autocorrelation (A)	18	1.12166057	147.72	< .0001	.65*
Model fitting (MF)	1	1.73754304	228.83	< .0001	.06*
MR*A	36	0.00037695	0.05	1.0000	< .06
A*MF	18	0.46427352	61.14	< .0001	.27*
MR*MF	2	0.02870814	3.78	.0323	< .06
Error	36	0.00759322			

Boldface β_1 (baseline slope), β_2 (level shift), and β_3 (slope change) denote that unacceptable estimates were uncovered under some manipulated conditions. * $p < .05$ and $\eta^2 \geq .06$

autocorrelation indicated that autocorrelation substantially impacted RMSEs of the β s, as compared to the effects of missing rate or model fitting. The RMSEs of β_1 and β_3 were all acceptable; their graphs are presented in the supplemental materials. Some of the RMSEs of β_2 were unacceptable, depending on the magnitude of autocorrelation. Boldface font in Table 2 identifies significant results combined with unacceptable RMSEs.

Figure 2 displays the RMSEs of β_2 obtained from all conditions manipulated. According to Fig. 2, the patterns of RMSEs are parallel across different missing rates, due to the nonsignificant interaction of missing rate and autocorrelation. Furthermore, the RMSEs obtained from the two models did not show noticeable differences, confirming the nonsignificant effect of model fitting. For the three missing rates, the RMSEs were unacceptable from $\varphi_1 = .1$ to $.7$.

RBES Table 3 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RBES for β_1 , β_2 , and β_3 . The main effects of autocorrelation and model fitting, plus their interaction, yielded significant results for the three β s. Some of the RBESs of the β s were unacceptable, depending on autocorrelation and the model fitted. The conditions in which the unacceptable RBESs occurred were similar across the three β s. Boldface font in Table 3 identifies significant results combined with unacceptable RBESs. Figure 3 displays the RBESs of β_2 obtained from all conditions manipulated. Similar figures displaying the RBESs of β_1 and β_3 can be found in the supplemental materials.

According to Fig. 3, the patterns of RBESs of β_2 are not parallel across the two models, due to the significant interaction of autocorrelation with model fitting. When the missing rate = 0% in A and 10% in B and the time-series model was fitted, RBESs were unacceptable only when $\varphi_1 = -.9$. In contrast, when the simplified model was fitted at the same missing rate, the RBESs were unacceptable when $\varphi_1 \leq 0$ or $\varphi_1 \geq .5$. When the missing rate = 10% in A and 20% in B and the time-series model was fitted, the RBESs were unacceptable when $\varphi_1 = -.9$ or $-.8$ or when $\varphi_1 \geq 0$. When the simplified model was fitted at the same missing rate, the RBESs were unacceptable when $\varphi_1 \leq -.2$ or $\varphi_1 \geq .3$. When the missing rate = 20% in A and 30% in B and the time-series model was fitted, the RBESs were unacceptable from $\varphi_1 = -.9$ to $-.7$ or when $\varphi_1 \geq -.3$. By comparison, when the simplified model was fitted at the same missing rate, RBESs were unacceptable when $\varphi_1 \leq -.3$ or $\varphi_1 \geq .1$.

$\beta_1 = \beta_2 = \beta_3 = 0.1$ paired with a long intervention phase ($n_A = 10$, $n_B = 56$)

RB Table 4 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RB for β_1 , β_2 , and β_3 . Similar to the results obtained from the short intervention phase, the main effects of missing rate and autocorrelation and their interaction yielded significant results for the three β s. Model fitting did not significantly affect the RBs of the three β s. The RBs of β_1 and β_3 were all acceptable; their graphs are presented in the supplemental materials. Some of the RBs of β_2 were unacceptable, depending on the missing rate and the magnitude of autocorrelation. Boldface font in Table 4 identifies significant results combined with unacceptable RB.

Figure 4 displays the RBs of β_2 obtained from all conditions manipulated. According to Fig. 4, the patterns of RBs are not parallel across different missing rates, due to the significant interaction of missing rates with autocorrelation. The RBs obtained from the two models did not

Fig. 6 RBESes of β_2 (level shift), when $\beta_1 = \beta_2 = \beta_3 = .1$ ($n_A = 10$, $n_B = 56$). SM stands for the simplified model, and TS stands for the lag-1 time-series model. Unacceptable RBESes are those below the bold reference line at RBESE = -0.1 or above the bold reference line at RBESE = 0.1

show noticeable differences, confirming the nonsignificant effect of model fitting. Specifically, when the missing rate = 0% in A and 10% in B, RBs were unacceptable from $\phi_1 = -.8$ to $.9$. When the missing rate = 10% in A and 20% in B, the RBs were unacceptable for all ϕ_1 s. When the missing rate = 20% in A and 30% in B, the RBs were acceptable for all ϕ_1 s.

RMSE Table 5 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RMSE for β_1 , β_2 , and β_3 . The main effects of missing rate and autocorrelation yielded significant results for the three β s. None of the two-way interactions were significant. Similar to the results obtained from the short intervention phase, the large η^2 s ($\geq .87$) due to autocorrelation indicated that autocorrelation substantially impacted the RMSEs of the β s, as compared with the effect of missing rate or model fitting. The RMSEs of β_1 and β_3 were all acceptable;

their graphs are presented in the supplemental materials. Some of the RMSEs of β_2 were unacceptable, depending on the magnitude of autocorrelation. Boldface font in Table 5 identifies significant results combined with unacceptable RMSEs.

Figure 5 displays the RMSEs of β_2 obtained from all conditions manipulated. According to Fig. 5, the patterns of RMSEs are parallel across different missing rates, due to the nonsignificant interaction of missing rate and autocorrelation. The RMSEs were unacceptable when $\phi_1 \geq .5$.

RBES Table 6 presents the ANOVA results for missing rate, autocorrelation, model fitting, and their two-way interactions on RBES for β_1 , β_2 , and β_3 . Similar to the results obtained from the short intervention phase, the main effects of autocorrelation and model fitting, plus their interaction, yielded significant results for the three β s. Some of the RBESes of β s were unacceptable, depending on autocorrelation and the model fitted. The

Table 7 Significant and (un) acceptable RB, RMSE, and RBESE

β_s	RB				RMSE								RBESE											
	$n_B = 10$				$n_B = 56$				$n_B = 10$				$n_B = 56$				$n_B = 10$				$n_B = 56$			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
β_1																								
MR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								
A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	×
MF																	×	×	×	×	×	×	×	×
MR*A	✓	✓	✓	✓	✓	✓	✓	✓																
A*MF																	×	×	×	×	×	×	×	×
MR*MF																								
β_2																								
MR	×	×	✓	×	×	×	✓	×	×	×	×	×	×	×	×	×								
A	×	×	✓	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
MF																	×	×	×	×	×	×	×	×
MR*A	×	×	✓	×	×	×	✓	×																
A*MF																	×	×	×	×	×	×	×	×
MR*MF																								
β_3																								
MR	✓	✓	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓								
A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	×
MF																	×	×	×	×	×	×	×	×
MR*A	✓	✓	✓	✓	✓	✓	✓	✓																
A*MF																	×	×	×	×	×	×	×	×
MR*MF																								

MR = missing rate, A = autocorrelation, and MF = model fitting. β s of 1 indicates all β s = .1, 2 indicates β_1 (baseline slope) = 1, 3 indicates β_2 (level shift) = 1, and 4 indicates β_3 (slope change) = 1. A “✓” denotes that the ANOVA test result was significant and the corresponding RB, RMSE, or RBESE was acceptable under all manipulated conditions. An “×” denotes that the ANOVA test result was significant but the corresponding RB, RMSE, or RBESE was unacceptable under some manipulated conditions. A blank entry indicates that the ANOVA test result was nonsignificant.

conditions for the unacceptable RBESEs were similar across the three β s. Boldface font in Table 6 identifies significant results combined with unacceptable RBESEs. Figure 6 displays RBESEs of β_2 obtained from all conditions manipulated. Similar figures displaying the RBESEs of β_1 and β_3 can be found in the supplemental materials.

According to Fig. 6, the patterns of the RBESEs of β_2 are not parallel across the range of autocorrelations between the two models fitted, confirming the significant interaction of autocorrelation with model fitting. When the missing rate = 0% in A and 10% in B and the time-series model was fitted, the RBESEs were unacceptable when $\varphi_1 = -.9$ or $-.8$. In contrast, when the simplified model was fitted at the same missing rate, the RBESEs were unacceptable when $\varphi_1 \leq -.2$ or $\varphi_1 \geq .2$. When the missing rate = 10% in A and 20% in B and the time-series

model was fitted, the RBESEs were unacceptable from $\varphi_1 = -.9$ to $-.7$ or from $\varphi_1 = -.1$ to $.6$. When the simplified model was fitted at the same missing rate, the RBESEs were unacceptable when $\varphi_1 \leq -.3$ or $\varphi_1 \geq .1$. When the missing rate = 20% in A and 30% in B and the time-series model was fitted, the RBESEs were unacceptable from $\varphi_1 = -.9$ to $-.7$ or from $\varphi_1 = -.2$ to $.7$. In contrast, when the simplified model was fitted at the same missing rate, the RBESEs were unacceptable when $\varphi_1 \leq -.3$ or $\varphi_1 \geq 0$.

Other magnitudes of β s paired with short and long intervention phases

In addition to the condition of $\beta_1 = \beta_2 = \beta_3 = .1$, three other specifications of the β s were investigated: (1) $\beta_1 = 1$ (increased baseline slope) while $\beta_2 = \beta_3 = .1$, (2) $\beta_2 = 1$ (increased level shift) while $\beta_1 = \beta_3 = .1$, and (3) $\beta_3 = 1$

Fig. 7 Applying EM to treat missing data in AB designs with a short intervention phase. The low missing rate = 0% in the A phase and 10% in the B phase. The medium missing rate = 10% in A and 20% in B, and the high missing rate = 20% in A and 30% in B

Fig. 8 Applying EM to treat missing data in AB designs with a long intervention phase. The low missing rate = 0% in the A phase and 10% in the B phase. The medium missing rate = 10% in A and 20% in B, and the high missing rate = 20% in A and 30% in B

(increased slope change) while $\beta_1 = \beta_2 = .1$. These specifications of the β s were investigated in short and long B phases. The results obtained from these specifications of β s were compared to those obtained from $\beta_1 = \beta_2 = \beta_3 = .1$ in short and long B phases. The findings revealed that changes in the β s did not affect the ANOVA results for RB, RMSE, and RBESE of the estimated β_1 , β_2 , and β_3 . However, when the results were assessed in terms of acceptability, the RBs of β_2 were all acceptable under Condition (2) only, even though the bias of β_2 remained unchanged regardless of β 's magnitude. The RBs of β_1 and β_3 were all acceptable under Conditions (1)–(3).

The RMSEs and RBESEs obtained from Conditions (1)–(3) were similarly acceptable or unacceptable, as compared to their counterparts obtained from $\beta_1 = \beta_2 = \beta_3 = .1$ with the identical missing rate, autocorrelation, intervention phase length, and model fitted.

Conclusions

On the basis of significance tests and the assessments of RB, RMSE, and RBESE in terms of acceptability, we concluded that EM was adequate as a missing-data method for AB designs. Table 7 summarizes the findings from all manipulated

conditions. The “✓” marks in Table 7 denote that the test result was significant and the corresponding criterion was acceptable under all manipulated conditions. An “×” denotes that the test result was significant but the corresponding criterion was unacceptable under some of the manipulated conditions. According to Table 7, missing rate, autocorrelation, and their interaction all had significant impacts on the RBs of β_1 , β_2 , and β_3 , regardless of intervention phase length or the magnitudes of the β s. Furthermore, the RBs of β_1 and β_3 were acceptable across all manipulated conditions. The RBs of β_2 were acceptable only when $\beta_2 = 1$ and $\beta_1 = \beta_3 = .1$, though the magnitude of bias remained the same under all magnitudes of the β s while holding other factors constant. Regardless of the intervention phase length or magnitudes of the β s, autocorrelation, with $\eta^2 \geq .86$, was the primary factor that significantly impacted the RMSEs of β_1 , β_2 , and β_3 . Missing rate, with $\eta^2 \leq .11$, secondarily impacted the RMSEs of β_1 , β_2 , and β_3 (only when the intervention phase was long). The RMSEs of β_1 and β_3 were acceptable across all manipulated conditions.

Autocorrelation, model fitting, and their interaction all had significant impacts on the RBESEs of β_1 , β_2 , and β_3 , regardless of intervention phase length or magnitudes of the β s. None of the RBESEs of the three β s were acceptable under all manipulated conditions.

Discussion

In the present study we sought to extend the findings of Smith et al. (2012) and Velicer and Colby (2005b, Study 2) on the performance of EM in treating missing data in an AB design. Specifically, this study investigated how missing-data rate, autocorrelation, intervention phase length, magnitude of the three β s (i.e., baseline slope, level shift, and slope change), and model fitting impacted the performance of EM. Three criteria (RB, RMSE, and RBESE) were employed to assess the quality of the β s' estimates as indicators of EM's performance. After data were simulated from a lag-1 interrupted time-series model, some were made missing under the MAR condition according to one of three missing rates: 0% in the A phase and 10% in the B phase (low missing rate), 10% in A and 20% in B (medium missing rate), and 20% in A and 30% in B (high missing rate). The performance of EM was evaluated in an ANOVA framework using each of the three criteria as a dependent variable. Significant results were identified by the combination of statistical significance level (p) < .05 and a medium $\eta^2 \geq .06$. The impact of the five factors on the performance of EM was further examined in terms of (un) acceptable RB, RMSE, and RBESE of the three β s. The RBs and RMSEs of β_1 and β_3 were acceptable under all manipulated

conditions. Table 7 summarizes both the significant and (un) acceptable results under all manipulated conditions.

The findings from our study revealed that, among the five factors manipulated, autocorrelation was the sole factor that impacted all three criteria singularly, especially the RMSEs of the estimated effects. The impact of autocorrelation on the RBs of the estimated effects interacted with missing rate. The impact of autocorrelation on the RBESEs depended on the model fitted. The interaction of autocorrelation with model fitting was not investigated in either Smith et al. (2012) or Velicer and Colby (2005b, Study 2). The interaction of autocorrelation with missing rate was examined descriptively by Smith et al.; they reported EM's worsening performance in power sensitivity as autocorrelation increased from 0 to .8. Velicer and Colby (2005b, Study 2) concluded that autocorrelation did not impact estimation of the effects, based on one positive and one negative autocorrelation. In comparison, our findings were based on 19 autocorrelations, ranging from extremely negative (−.9) to extremely positive (.9). This wide range of autocorrelation enabled us to comprehensively examine the impact of autocorrelation, in terms of both magnitude and direction, on three qualities of the estimated effects, especially the RMSE.

Missing-data rate impacted the RBs and RMSEs of the estimated effects singularly. As compared to autocorrelation, the impact of missing rate on RMSEs was much smaller. Missing rate did not impact RBESEs. These findings partially support the results from Smith et al. (2012) and Velicer and Colby (2005b, Study 2). Smith et al. reported that missing proportion did not impact the power sensitivity of the level-change effects, and Velicer and Colby (2005b, Study 2) reported that missing rate did not impact the maximum-likelihood estimation of the effects. It is worth noting that the implementation of missing-data rate in the present study ensured a lower missing rate in the A phase than in the B phase. In contrast, Smith et al. implemented a missing rate for the A and B phases combined, whereas Velicer and Colby (2005b, Study 2) implemented equal missing rate for the A and B phases.

The two models fitted to the completed data did significantly impact the RBESEs of the estimated effects, but not the RBs or RMSEs. Our results from fitting a simplified model to completed data are consistent with Harrington and Velicer (2015), who reported biased estimates of the error variance if autocorrelation was not accounted for. Specifically, negative autocorrelations that were unaccounted for resulted in overestimation of the error variance, and positive autocorrelations unaccounted for resulted in underestimation of the error variance (Harrington & Velicer, 2015). Our findings also revealed encouraging results obtained from fitting a simplified model—that is, when baseline slope and slope change were used as the indicators of an intervention effect.

The effect of intervention phase length on the performance of EM was investigated in the present study by manipulating the B phase length to be either short (ten scores) or long (56 scores). The findings from our study revealed that the two intervention phase lengths yielded similar results when missing-data rate, autocorrelation, and model fitted were held constant. These results confirmed that the estimation of the effects was not sensitive to the intervention phase length under the manipulated conditions. Because our findings uncovered a major impact due to autocorrelation on the RBs, RMSEs, and RBESEs of estimated effects, researchers are advised to estimate autocorrelation correctly before using EM to treat missing data, unless the baseline slope and slope change are used as joint indicators of an intervention effect, as noted above. The literature has recommended at least 20 data points per phase in order to obtain stable estimates of autocorrelation (Box & Jenkins, 1976; Crosbie, 1993; Velicer & Colby, 2005a). To estimate autocorrelation reliably with a short baseline or intervention phase, readers should consult Shadish, Rindskopf, Hedges, and Sullivan (2013).

The magnitudes of three effects did not impact RBs, RMSEs, or RBESEs. In fact, the impact of changes in β s on these three criteria can be derived analytically (see the supplemental materials for the proof). The results from the present study provided empirical evidence of the impact. An increase in β_1 , β_2 , or β_3 (i.e., from .1 to 1) simply adds a constant to the estimated effect and its corresponding true value, resulting in the same magnitude of bias. Therefore, RB is adjusted accordingly by definition. RMSE and RBESE should remain the same, regardless of a change in the β s, because bias and sampling variance remained the same in this simulation study. Solanas et al. (2010) reported the same findings when slope change and level change were estimated separately, using their new procedure that controlled for a linear baseline trend.

On the basis of all findings uncovered in the present study, we constructed two decision trees (Figs. 7 and 8), one for each intervention phase length, to assist researchers and practitioners in applying EM to deal with missing data. The decision trees separate inference of the baseline slope and slope change from that of level shift. Under each inferential purpose, we further ask whether population effects are to be tested statistically. If a statistical test is indeed to be performed on the estimated baseline slope and slope change, or level shift, then the selection of a model (time-series vs. simplified) is relevant. Otherwise, model selection is irrelevant. At the end of each decision tree, we offer our recommendations on the conditions under which EM's treatment of missing data is acceptable in terms of RB, RMSE, and RBESE.

Limitations and suggestions for future research

We designed the present study to improve on Smith et al. (2012) and Velicer and Colby (2005b, Study 2) in nine

respects: MAR condition for missing data, differential missing rates for A and B phases, two intervention phase lengths, a wide range of autocorrelations, two models fitted to the imputed data, three effects in AB designs, four magnitudes of the three effects, a large number of replications, and three criteria to gauge the performance of EM. Our findings summarized in Table 7 and Figs. 7 and 8 provide empirical evidence supporting EM as an advanced missing-data method for AB designs. As we previously stated, EM is readily available in several computing tools, and its programming is easy. EM should be recommended as a viable option in dealing with missing data in SCED studies.

Our study is nonetheless limited by its scope. In the present study we applied simplified and time-series models to fit the completed data; both assumed that the data were gathered at equally spaced intervals. The impact of violating this equal-interval assumption will require additional studies. In the present study we constrained the error variance to a constant at each time point according to suggestions in the literature. Future research could examine the performance of EM or other modern missing-data methods without this constraint, so as to mimic data that exhibit different variabilities in different phases or at different time points. The curvilinear relationship uncovered in this study between RMSE and autocorrelation requires further investigation in order to fully understand the rationale behind this nonlinear pattern. Our study revealed an encouraging possibility of fitting time-series data with a simplified model, without autocorrelation, when baseline slope and slope change were the indicators of an intervention effect. Additional studies will be needed to further investigate whether the simplified model can be used with other types of SCEDs, such as alternating-treatment or multiple-baseline designs. Finally, although the present study extended Smith et al. (2012) and Velicer and Colby (2005b, Study 2) by studying the impact of the intervention phase length on the performance of EM, future research might systematically examine the impact of the baseline phase length on the performance of EM.

SCED studies have been paramount in establishing and confirming evidence-based practices (Horner et al., 2005; Shadish & Sullivan, 2011; Smith, 2012). It is therefore imperative that SCED research be conducted at the highest level of rigor, to yield credible and generalizable results. Missing SCED data should not be ignored in visual or statistical inferences; they should be properly treated. It is our hope that the findings from this study of the EM algorithm will contribute to the current scholarship on SCED methodology and provide a sense of emerging research directions for missing-data treatment in SCED studies.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Allison, D., & Gorman, B. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior Research and Therapy*, 31, 621–631. [https://doi.org/10.1016/0005-7967\(93\)90115-B](https://doi.org/10.1016/0005-7967(93)90115-B)
- Allison, D. B., Silverstein, J. M., & Gorman, B. S. (1996). Power, sample size estimation, and early stopping rules. In R. D. Franklin, D. B. Allison, & B. S. Gorman, (Eds.), *Design and analysis of single-case research* (pp. 335–371). Mahwah: Erlbaum.
- Barnard, J. (2000). Multiple imputation for missing data. Paper presented at the Summer Workshop of the Northeastern Illinois Chapter of the American Statistical Association, Northbrook.
- Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.
- Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification*, 39, 835–869. <https://doi.org/10.1177/0145445515603706>
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Couvreux, C. (1997). The EM algorithm: A guided tour. In M. Kárný & K. Warwick (Eds.), *Computer intensive methods in control and signal processing* (pp. 209–222). Boston: Birkhäuser. https://doi.org/10.1007/978-1-4612-1996-5_12
- Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, 61, 966–974. <https://doi.org/10.1037/0022-006X.61.6.966>
- 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*, 39, 1–38.
- Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2, 222. <https://doi.org/10.1186/2193-1801-2-222>
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods*, 42, 930–943. <https://doi.org/10.3758/BRM.42.4.930>
- Gustafson, S. A., Nassar, S. L., & Waddell, D. E. (2011). Single-case design in psychophysiological research. Part I: Context, structure, and techniques. *Journal of Neurotherapy*, 15, 18–34. <https://doi.org/10.1080/10874208.2011.545762>
- Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, 50, 162–183. <https://doi.org/10.1080/00273171.2014.973989>
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, 26, 329–367. <https://doi.org/10.1177/0049124198026003003>
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165–179. <https://doi.org/10.1177/001440290507100203>
- Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, 60, 38–58. <https://doi.org/10.1177/00131640021970358>
- Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods*, 39, 343–349. <https://doi.org/10.3758/BF03193002>
- Kratochwill, T. R. (2015). Single-case research design and analysis: An overview. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 1–14). Mahwah: Erlbaum.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd). New York: Wiley.
- Manolov, R., & Solanas, A. (2008). Comparing $N = 1$ effect size indices in presence of autocorrelation. *Behavior Modification*, 32, 860–875. <https://doi.org/10.1177/0145445508318866>
- Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for qualifying single-case intervention effectiveness. *Behavior Therapy*, 42, 533–545. <https://doi.org/10.1016/j.beth.2010.12.003>
- Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17, 372–411. <https://doi.org/10.1177/1094428114548590>
- Parker, R. I., Brossart, D. R., 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*, 205, 116–132.
- Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Charlotte: Information Age.
- Richards, S. B., Taylor, R. L., & Ramasamy, R. (2014). *Single subject research: Applications in educational and clinical settings* (2nd). Belmont: Wadsworth.
- SAS Institute Inc. (2015). *SAS/STAT 14.1 user's guide*. Cary: Author.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schlomer, G. L., Bauman, S., & Card, N. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, 57, 1–10. <https://doi.org/10.1037/a0018082>
- Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, 43, 813–821. <https://doi.org/10.3758/s13428-012-0282-1>
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, 43, 971–980. <https://doi.org/10.3758/s13428-011-0111-y>
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17, 510–550. [https://doi.org/10.1016/S0149-2063\(99\)80083-X](https://doi.org/10.1016/S0149-2063(99)80083-X)
- 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, 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, 195–218. <https://doi.org/10.1177/0145445510363306>
- Springer, T., & Urban, K. (2014). Comparison of the EM algorithm and alternatives. *Numerical Algorithm* 67(2), 335–364. <https://doi.org/10.1007/s11075-013-9794-8>

- SPSS Inc. (2010). PASW Statistics for Windows and Mac (Version 18.0.0). Chicago: Author.
- Velicer, W. F., & Colby, S. M. (2005a). A comparison of missing-data procedures for ARIMA time-series analysis. *Educational and Psychological Measurement*, 65, 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). Mahwah: Erlbaum.
- Velicer, W. F., & Fava, J. (2003). Time series analysis. In J. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 581–606). New York: Wiley.
- What Works Clearinghouse. (2017). *Standards handbook* (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf
- Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook, T. D. (2012). Research designs for program evaluation. In J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (2nd, pp. 316–341). Hoboken: Wiley.