Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect

Avi Kenny[*,1], **Emily Voldal**[1], **Fan Xia**[1], **Patrick J. Heagerty**[1], **James P. Hughes**[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington

Summary

Stepped wedge cluster randomized controlled trials are typically analyzed using models that assume the full effect of the treatment is achieved instantaneously. We provide an analytical framework for scenarios in which the treatment effect varies as a function of exposure time (time since the start of treatment) and define the "effect curve" as the magnitude of the treatment effect on the linear predictor scale as a function of exposure time. The "time-averaged treatment effect", (TATE) and "long-term treatment effect" (LTE) are summaries of this curve. We analytically derive the expectation of the estimator $\hat{\delta}$ resulting from a model that assumes an immediate treatment effect and show that it can be expressed as a weighted sum of the time-specific treatment effects corresponding to the observed exposure times. Surprisingly, although the weights sum to one, some of the weights can be negative. This implies that $\hat{\delta}$ may be severely misleading and can even converge to a value of the opposite sign of the true TATE or LTE. We describe several models, some of which make assumptions about the shape of the effect curve, that can be used to simultaneously estimate the entire effect curve, the TATE, and the LTE. We evaluate these models in a simulation study to examine the operating characteristics of the resulting estimators and apply them to two real datasets.

Keywords

stepped wedge; cluster randomized trial; time-varying treatment effect; model misspecification

1 | INTRODUCTION

Cluster randomized trials (CRTs) involve randomizing groups of individuals to a treatment or control condition, and are often conducted when individual randomization is impractical. Cluster randomized designs are often adopted to evaluate the impact of health care system level interventions through modification of the systematic processes used to treat patients. One type of CRT design is the stepped wedge, which has seen increased usage in recent

years.[1] In a stepped wedge CRT, all clusters begin in the control state and eventually switch over to the treatment state in a staggered manner, with a random assignment of clusters to crossover times, or "sequences". Data are typically collected from all clusters at each time point, often through a series of cross-sectional surveys. Some unique aspects of this design include the partial confounding of the treatment effect with time and the fact that all clusters are observed in both the control state and the treatment state. The strengths and limitations of the stepped wedge design have been extensively discussed in recent years.[2,3,4,5,6,7,8] In particular, the stepped wedge design is useful for situations in which it is not possible to implement the treatment simultaneously to all participants for logistical, financial, or other reasons.[9]

Standard statistical models for analyzing data from stepped wedge CRTs typically include a treatment indicator variable, indexed by cluster and time, that equals zero when the cluster is in the control state and one when the cluster is in the treatment state.[10] The coefficient of this indicator variable can then interpreted as the treatment effect. This modeling choice implicitly assumes that the full effect of the treatment is reached immediately (i.e. within a single time step) and does not increase or decrease thereafter; for this reason, we refer to such models as *immediate treatment effect* (IT) models. Thus, the IT model assumes that the shape of the *effect curve* – the level of the treatment effect as a function of time since the start of the treatment – is fully known; this shape is depicted in Figure 1a. However, with the stepped wedge design, once each cluster crosses over to the treatment condition it will experience multiple time periods under the treatment condition with the possibility that the magnitude of treatment effect will change with increasing treatment experience. Such effect modification with increasing time is often referred to as the "learning curve" and is well-documented in areas such as surgery.[11] As such, the assumption of an immediate treatment effect may be violated in some settings.

We follow Nickless et al.[12] and use the term "exposure time" to refer to the amount of time that has passed since the start of the treatment for a given cluster, where the start of the treatment corresponds to exposure time 0. This is contrasted with "study time", which is the amount of time that has passed since the start of the study. There are several ways in which a treatment effect can vary with exposure time. The effect might not be realized until the second or third time point following the treatment, but then reaches its full effect almost immediately; we refer to this as a "delayed effect" (Figure 1b). Alternatively, the magnitude of the effect might vary as a function of exposure time (Figure 1c,d). It is also possible for the treatment effect to vary as a function of study time (e.g. if an external event affects the intervention across the entire study population at once), but we do not consider this case in this paper.

In some scenarios, the assumption of an immediate treatment effect seems justifiable. For example, it is reasonable to think that the effect of the implementation of a surgical safety checklist on patient outcomes would be immediate.[13] However, this assumption feels less plausible in many other scenarios. For example, the effect of education and counseling programs on exclusive breastfeeding rates may take months or years to achieve, since behavior change can be a slow process.[14] As another example, community health programs

often involve a series of training modules for health workers that occur over a period of months, each of which is designed to have a separate effect on under-five mortality.[15]

The problem of a time-varying treatment effect in the context of stepped wedge CRTs was first considered by Granston et al.[16], who demonstrated that modeling the treatment effect using an indicator function can lead to biased estimates and invalid inference when the true effect curve is time-varying. They introduced a parametric model to account for the time-varying effect which assumes that the effect curve falls in a two-parameter family of concave functions. However, this model involves a complex two-stage estimation process and leads to incorrect inference when the true effect curve does not fall within this family, and fails to converge when the true effect curve is at least partially convex, as in Figure 1d (see supplementary material).

Hughes et al.[1] further considered the problem of a time-varying treatment effect and noted that in certain cases, this issue can be prevented by careful study design, such as increasing the length of time between steps, switching from an outcome endpoint that may take years to change to a process endpoint that will change more rapidly, or including a "washout period" immediately after implementation during which no data are collected. They also suggested the approach of changing the indicator variable representing the presence of the treatment such that this variable can take on fixed values between zero and one. However, this assumes that the shape of the effect curve is fully known.

Hemming et al.[17] described a model that includes fixed effects corresponding to study time by treatment interaction terms (as opposed to exposure time by treatment interactions), with study time treated as a categorical variable. They concluded that the confidence intervals for each interaction term coefficient were too wide for the model to be useful; however, they did not consider combining the interaction term coefficients into a single estimator.

As part of a simulation study, Nickless et al.[12] tested several models that account for time-varying treatment effects. They consider four types of interaction terms between time and treatment, which differ in terms of how time is modeled. Time can be either study time or exposure time, and can be modeled as continuous or categorical. The model that includes an interaction between treatment and categorical study time is equivalent to the Hemming et al.[17] model.

Although several authors have touched on the issue of time-varying treatment effects, we see several major gaps in the literature. First, no one has characterized the behavior of the standard treatment effect estimator when the assumption of an immediate treatment effect is violated. Second, there is ambiguity and lack of precise terminology around estimands of interest. Third, different models that account for time-varying treatment effects have not been studied in a unified manner.

This paper is organized as follows. In section 2, we analytically examine the behavior of the class of models that assume an immediate treatment effect and show that they may exhibit counterintuitive behavior in the presence of time-varying treatment effects. In section 3, we define several potential estimands of interest, introduce several models that account for time-varying treatment effects, and discuss how these models can be used for estimation and

hypothesis testing in the context of both confirmatory and exploratory analyses. In section 4, we perform a simulation study to test the behavior of the models across a selection of data-generating mechanisms. In section 5, we illustrate the use of these models in real datasets. In section 6, we discuss these results and their implications for the analysis of stepped wedge trials.

2 | BEHAVIOR OF THE IT MODEL UNDER A TIME-VARYING TREATMENT EFFECT

In this section, we analytically examine the behavior of the IT model treatment effect estimator when the true treatment effect varies as a function of exposure time. Throughout this paper, we assume that the data come from a "standard" stepped-wedge trial, meaning that there are J equally-spaced time steps, $Q = J - 1$ sequences, and an equal number of clusters per sequence, with cross-sectional measurements taken at each time point. Assume the outcome data are continuous, and let Y_{ijk} denote the observed outcome for individual $k \in \{1, \ldots, K\}$ within cluster $i \in \{1, \ldots, I\}$ at time $j \in \{1, \ldots, J\}$. Also let sequences be labeled such that in sequence q, the treatment is introduced at time point $j = q + 1$. For simplicity, we focus our analysis on the Hussey and Hughes model[10], a special case of the IT model that is commonly used to analyze data from stepped wedge trials, although the behavior we observe holds more generally; we discuss this further at the end of this section. This model is specified in (1), where the β_j terms represent the underlying time trend (with $\beta_1 = 0$ for identifiability), $\alpha_1, \ldots, \alpha_I \overset{iid}{\sim} N(0, \tau^2)$ is a set of random effects accounting for the dependence of observations within a cluster, and $\epsilon_{111}, \ldots, \epsilon_{IJK} \overset{iid}{\sim} N(0, \sigma^2)$ are residual error terms.

$$Y_{ijk} = \mu + \beta_j + \delta X_{ij} + \alpha_i + \epsilon_{ijk}$$
$$X_{ij} = \begin{cases} 1, & \text{cluster } i \text{ is in the treatment state at time } j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Next, let s_{ij} represent the exposure time of cluster i at time j and let $\delta(s)$ represent the treatment effect at exposure time s, which we refer to as the "point treatment effect" (PTE) at s. Model (2) incorporates the time-varying treatment effect $\delta(s_{ij})$, where X_{ij} is defined as above:

$$Y_{ijk} = \mu + \beta_j + \delta(s_{ij})X_{ij} + \alpha_i + \epsilon_{ijk} \tag{2}$$

Suppose that data are generated according to (2) but analyzed with the model specified in (1). We are interested in how the estimator $\hat{\delta}$ behaves in this scenario. Since s_{ij} takes on values between 1 and $J - 1$, it would not be unreasonable to expect the IT estimator to converge to a value close to the average treatment effect over the exposure period, lying between the smallest and largest of the time-specific treatment effects $\delta(1), \ldots, \delta(J - 1)$. However, this turns out to not necessarily be the case. This conclusion follows from Theorem 1, in which we provide closed-form expressions for the treatment effect estimator resulting from model (1) and its expectation.

Theorem 1. Suppose we have a standard stepped wedge design with data generated according to (2), which may involve multiple clusters per sequence. If we fix $\phi \equiv \tau^2/(\tau^2 + \sigma^2/K)$ and denote the mean outcome in sequence $q \in \{1, \ldots, Q\}$ at time point j by $\bar{Y}_{qj}$, the treatment effect estimator $\hat{\delta}_\phi$ obtained by fitting model (1) via weighted least squares can be expressed as:

$$\hat{\delta}_\phi = \frac{12(1+\phi Q)}{Q(Q+1)(\phi Q^2 + 2Q - \phi Q - 2)} \sum_{q=1}^{Q} \sum_{j=1}^{J}$$
$$\left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1+\phi Q)} \right] \bar{Y}_{qj} \tag{3}$$

Furthermore, the expectation of the treatment effect estimator $\hat{\delta}_\phi$ can be expressed as a weighted average of the time-specific treatment effects:

$$E[\hat{\delta}_\phi] = \sum_{s=1}^{J-1} w(Q, \phi, s)\delta(s), \tag{4}$$

where the weights are defined as follows:

$$w(Q, \phi, s) \equiv \frac{6(s - Q - 1)((1 + 2\phi Q)s - (1 + \phi + \phi Q)Q)}{Q(Q+1)(\phi Q^2 + 2Q - \phi Q - 2)} \tag{5}$$

Proof. See Appendix D.

The expectation $E[\hat{\delta}_\phi]$ is thus a weighted sum of the individual time-specific treatment effects $\delta(1), \ldots, \delta(J-1)$, and does not depend on the underlying study time trend. It is easily shown that $\sum_{s=1}^{J-1} w(Q, \phi, s) = 1$ for any pair (Q, ϕ), which is necessarily the case since otherwise $\hat{\delta}$ would be biased when the IT model is correct. However, surprisingly and unfortunately, *some of the weights can be negative*, depending on the values of ϕ and Q. Additionally, a simple calculation shows that $w(Q, \phi, J-1)$, the weight corresponding to $\delta(J-1)$, will *always* be negative (assuming $\phi > 0$). Equation (5) is somewhat difficult to interpret in itself; the impact of this finding is perhaps best illustrated in Figure 2, which shows the expected effect curve estimated from model (1) – which is necessarily constant after the first exposure time point – plotted against the true curve for four possible effect curves.

Looking at panel (a) of Figure (2), we see that when the assumption of an immediate treatment effect is correct, we correctly estimate the effect curve; this is expected, as weighted least squares estimators in linear mixed models are unbiased in general. However, the results from panels (b), (c), and (d) show that when the immediate treatment effect assumption is violated, the estimated effect curve can be astonishingly misleading. In panels (b) and (d), the estimated effect curve lies entirely below the true effect curve. Furthermore, we observe that even if each time-specific treatment effect is positive (or zero), the estimated treatment effect can actually converge to a negative value! This in turn implies that, if we

are hoping to estimate the average treatment effect over the exposure period or the long-term effect, our estimator can converge to a value of the opposite sign as the true value.

The value $\phi = 0.5$ used to compute the estimates in Figure (2) is not unreasonable for a stepped wedge design. As noted by Matthews and Forbes[18], large numbers of individuals per cluster can lead to large values of ϕ, even if the intraclass correlation coefficient (ICC) $\rho \equiv \tau^2/(\tau^2 + \sigma^2)$ is small. For example, for an ICC of 0.05, $n = 10$ leads to $\phi = 0.34$ and $n = 50$ leads to $\phi = 0.72$. In Appendix A, we provide a figure showing how the weights vary as a function of ϕ for select values of Q, as implied by (5). Additionally, we note that ϕ is generally not known and must be estimated from the trial data, and so the estimator we actually use is $\hat{\delta} \equiv \hat{\delta}_{\hat{\phi}}$. Since the model is misspecified when the treatment effect varies with time, the variance component estimators will not, in general, be consistent for the true values.

We also note that this counterintuitive behavior is not restricted to model (1). In Appendix A, we conduct a numerical analysis to compute the weights corresponding to a model that includes the addition of a random time effect (i.e. a random intercept corresponding to a cluster-by-time interaction term, as in the Hooper/Girling model[19]), such that the correlation matrix has a nested exchangeable structure. [20] We also perform the same analysis for a model involving a random treatment effect. For both alternative correlation structures, the resulting weights are qualitatively similar to those resulting from model (1), in the sense that the largest weight is on $\delta(1)$ and some weights are negative. This analysis shows that the counterintuitive behavior of the IT model treatment effect estimator occurs across a wide range of models that are commonly used for analyzing stepped wedge trials.

Furthermore, this issue will also apply to models fit using generalized estimating equations (GEE). If a researcher fits a GEE model using the fixed effects structure of (1) and a working exchangeable correlation structure, the results of Theorem 1 will still apply; that is, the resulting treatment effect estimator will still converge to a counterintuitive value. If instead, a working independence correlation structure is used, then the treatment effect estimator may converge to a counterintuitive value as well, but a different counterintuitive value. Specifically, instead of result (4), we will have $E[\hat{\delta}] = \sum_{s=1}^{J-1} w(Q, 0, s)\delta(s)$. This is not as bad in the sense that $w(Q, 0, s) \geq 0$ for all Q, s, but it still leads to a poorly-defined estimand.

3 | METHODS

As demonstrated in section (2), a model that assumes an immediate treatment effect can lead to severely misleading inference if the treatment effect varies with exposure time. This necessitates the development of models that explicitly account for time-varying treatment effects. However, in the absence of an immediate and constant treatment effect, the estimand of interest is less well-defined. Therefore, before presenting these models, we define our estimands of interest more formally.

3.1 | Estimands of interest

In this section, we formally introduce a model for the data-generating mechanism for the purpose of defining estimands of interest. Consider a standard stepped wedge design, as defined in section 2. Let $Y_{ijk}(q)$ denote the potential outcome for individual $k \in \{1, \ldots, K\}$ within cluster $i \in \{1, \ldots, I\}$ at time $j \in \{1, \ldots, J\}$, had this cluster been assigned to sequence $q \in \{1, \ldots, J-1\}$, where sequences are labeled such that in sequence q, the treatment is introduced at time point $j = q + 1$. Also let $\mu_{ijk}(q)$ denote the expectation of $Y_{ijk}(q)$. We assume the following structural model for $\mu_{ijk}(q)$, an adaptation of the generic structural mixed model of Li et al.[20]:

$$g\big(\mu_{ijk}(q)\big) = \overbrace{\Gamma_j}^{\text{time trend}} + \overbrace{\Delta_{ijq}}^{\text{treatment effect}} + \overbrace{C_{ijkq}}^{\text{heterogeneity}} \tag{6}$$

Above, $g(\cdot)$ is a link function, and the outcome is assumed to fall within a parametric family with mean $\mu_{ijk}(q)$. Γ_j, represents the underlying study time trend and Δ_{ijq} represents the treatment effect. The C_{ijkq} term accounts for the correlation structure of the data, and will often take the form of one or more mean-zero random effect terms to capture cluster-level or temporal deviations from the fixed effects.

Let the random variable Y_{ijk} denote the observed outcome. The observed design matrix after randomization is $X \equiv \{X_{ij} : i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}\}$, where $X_{ij} = 1$ if cluster i is is assigned to be in the treatment state at time j. If we assume that clusters are assigned to sequences randomly (exchangeability) and that $Y_{ijk}(q) = Y_{ijk}$ if cluster i is randomly assigned to sequence q (consistency), then the potential outcome expectation $\mu_{ijk}(q)$ is identified by $\mu_{ijk} \equiv E[Y_{ijk}|X_{ij}]$, and structural model (6) implies:

$$g\big(\mu_{ijk}\big) = \Gamma_j + \Delta_{ij}X_{ij} + C_{ijk} \tag{7}$$

In the setting of randomized trials, these assumptions will typically hold. In equation (7), the q subscripts have been removed since they are determined by i and j given X_{ij}. Other works have focused on choices for the time trend Γ_j[12,17,20] and for the heterogeneity component C_{ijk}.[20,21,22,19,23] In this paper, we restrict attention to modeling choices related to the treatment effect Δ_{ij}. Furthermore, we assume that the treatment effect may vary as a function of exposure time, but *not* as a function of study time. If s_{ij} again represents the exposure time of cluster i at time j, and δ is a fixed function, then this assumption allows us to simplify model (7) as follows:

$$g\big(\mu_{ijk}\big) = \Gamma_j + \delta(s_{ij})X_{ij} + C_{ijk} \tag{8}$$

This model makes no assumptions about the form of the time trend or the heterogeneity, and is thus general enough to apply in a number of settings, such as when there are random time effects or random treatment effects.

In this context, referring to a single parameter as the "treatment effect" is ambiguous, since the treatment effect varies as a function of exposure time. Therefore, we must state more precisely what we are interested in estimating. First, we define the *effect curve* as the function $s \mapsto \delta(s)$, where we think of time as continuous even though the analysis of stepped wedge data typically involves assigning observations to one of a discrete set of time points $\{1, \ldots, J\}$. This function is an appealing estimand since it contains a wealth of information that can be used to guide program design and evaluation. However, in many applications it will be desirable to focus on a single parameter that summarizes this curve. The first summary estimand we consider is the time-averaged treatment effect (TATE) from exposure time s_1 to exposure time s_2, which is the average value of the effect curve over the interval $(s_1, s_2]$. We denote this estimand by $\Psi_{(s_1, s_2]}$ and formally define it as follows:

$$\Psi_{(s_1, s_2]} \equiv \frac{1}{s_2 - s_1} \int_{s_1}^{s_2} \delta(s) ds \tag{9}$$

For a given study that involves J discrete time points, one choice of the interval $(s_1, s_2]$ will be $(0, J-1]$, the entire exposure period. Alternatively, a researcher may choose to increase the lower endpoint if he or she is interested in the TATE after a washout period immediately following implementation of the treatment. Care should be taken when comparing TATE estimates from two different studies, even if they both examine the same intervention and use the same outcome measure, since the TATE is defined relative to a time interval, and these time intervals may differ between studies.

The second summary estimand we consider is the point treatment effect (PTE) at exposure time s_0, which is simply defined as the value $\Psi_{s_0} \equiv \delta(s_0)$. A common choice for s_0 might be $J-1$, the maximum treatment period. If we additionally assume that the effect curve "flattens out" by the end of the study (that is, we have that $\delta(s) = \delta(J-1)$ whenever $s \quad J-1$), Ψ_{J-1} can be interpreted as the long-term treatment effect (LTE).

When the true effect curve has no delay or change over time (i.e. $\delta(s)$ is constant), we have that $\Psi_{(s_1, s_2]} = \Psi_{s_0}$ for all s_0, s_1, s_2. However, in the presence of a time-varying effect, these estimands do not in general coincide, and so in the context of a confirmatory trial, the researcher must specify one in advance as the primary endpoint. Others can be examined within secondary or exploratory analyses, if applicable. We expect that researchers will most often be interested in a TATE or the LTE as the target for confirmatory trials, whereas estimation of the effect curve $s \mapsto \delta(s)$ will usually be done as an exploratory analysis. Since different researchers will be interested in different estimands, we will consider estimation of all three in this paper.

3.2 | Analysis models

Model (8) was written generically so that the estimands are well-defined in a variety of settings. The analysis models we present in this section can be seen as special cases of (8) that make assumptions about the treatment effect term $\delta(s)X_{ij}$. Implementation of these models in specific settings will involve making context-appropriate choices for the

Γ_j and C_{ijk} terms. For example, we may choose to take $\Gamma_j = \mu + \beta_j$ and $C_{ijk} = \alpha_i$ with $\alpha_1, \ldots, \alpha_I \overset{iid}{\sim} N(0, \tau^2)$ to yield a class of models containing the Hussey and Hughes model.[10] All models in this section can be used for the purposes of estimation and hypothesis testing related to both the TATE and the PTE/LTE.

Immediate treatment effect (IT) model—First, we consider the *immediate treatment effect* (IT) model, a generalization of the model we studied in section 2. This model assumes that full effect is reached within a single time step and does not change thereafter, as in Figure 1a; we provide a specification of this model in equation (10), where δ is a scalar parameter. Most mixed models that have previously been used to analyze data from stepped wedge trials, such as the Hussey and Hughes model[10] and the Hooper/Girling model[19], are special cases of the IT model.

$$g(\mu_{ijk}) = \Gamma_j + \delta X_{ij} + C_{ijk}$$
$$X_{ij} = \begin{cases} 1, & \text{cluster } i \text{ is in the treatment state at time } j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

As noted, when the IT model is correct (i.e. there is no variation in the treatment effect over time), the TATE (for any (s_1, s_2)) and PTE (for any s_0) coincide, and so we can estimate both by fitting this model and using $\hat{\delta}$ as our estimator. When the IT model is correct, we will show that $\hat{\delta}$ is a more efficient estimator of the TATE or PTE than the other estimators we will consider. However, when the IT model is incorrect, $\hat{\delta}$ can give highly misleading results, as described in section 2.

Exposure time indicator (ETI) model—Next, we consider a model that makes no assumptions about the shape of the effect curve, which we refer to as the *exposure time indicator* (ETI) model. This is generalization of models considered in Granston et al.[16] and Nickless et al.[12] and is specified by equation (11), where, with a slight abuse of notation, we use the subscript s as shorthand for s_{ij}, the exposure time of cluster i at time j.

$$g(\mu_{ijk}) = \Gamma_j + \delta_s X_{ij} + C_{ijk} \tag{11}$$

Model (11) is saturated with respect to exposure time; as such, it is the most flexible model we consider. After fitting this model and obtaining parameter estimates $\hat{\delta}_1, \ldots, \hat{\delta}_{J-1}$, we can base estimation and inference on the following estimators, where we assume that s_0, s_1, and s_2 are all integers in the set $\{0, \ldots, J-1\}$ with $s_1 < s_2$.

$$\widehat{\Psi}_{(s_1, s_2]} \equiv \frac{1}{s_2 - s_1} \sum_{r=1}^{s_2 - s_1} \hat{\delta}_{s_1 + r}$$
$$\widehat{\Psi}_{s_0} \equiv \hat{\delta}_{s_0} \tag{12}$$

The estimator $\widehat{\Psi}_{(s_1, s_2]}$ can be viewed as a right-hand Riemann sum that approximates the integral $\Psi_{(s_1, s_2]}$. Alternatively, a researcher can use an estimator based on a trapezoidal

Riemann sum; this is discussed in Appendix C. The estimator $\widehat{\Psi}_{J-1}$ can be used to estimate the LTE if it is assumed to exist, although, as we will see, the variance of this estimator will typically be quite high. Given this, another possibility is to assume that the long-term effect is reached by some time point $J - m$ (with $m > 1$) before the end of the study and estimate the LTE using the TATE estimator $\widehat{\Psi}_{(J-m, J-1]}$ to gain precision.

A key assumption of this model is that the shape and height of the effect curve do not vary between clusters; this is a fairly strong assumption that we will later relax when we discuss random treatment effects in section 3.3. We also note that the IT model is a submodel of the ETI model, and so (likely in the context of an exploratory analysis), a researcher can conduct a likelihood ratio test to determine whether the IT model is an appropriate simplification of the ETI model for a given dataset.

Since $\widehat{\Psi}_{(s_1, s_2]}$ is a linear combination of the parameter estimates $\hat{\delta}_1, \ldots, \hat{\delta}_{J-1}$, variance estimation is straightforward. Assuming the $(J-1) \times 1$ matrix $\hat{\delta} \equiv (\hat{\delta}_1, \ldots, \hat{\delta}_{J-1})^T$ is approximately multivariate Normal with estimated covariance matrix $\hat{V}$, we find the $1 \times (J-1)$ matrix M such that $\widehat{\Psi}_{(s_1, s_2]} = M\hat{\delta}$. Our estimated variance is then $\widehat{\text{Var}}(\widehat{\Psi}_{(s_1, s_2]}) \equiv M\hat{V}M'$, where M' represents the transpose of M. For example, if we are estimating $\widehat{\Psi}_{(0, J-1]}$, we have that $M = \left(\frac{1}{J-1}, \ldots, \frac{1}{J-1}\right)$ and $M\hat{V}M' = \frac{1}{(J-1)^2} \sum_{i=1}^{J-1} \sum_{j=1}^{J-1} \hat{V}_{ij}$.

Natural cubic spline (NCS) model—If we are willing to assume some degree of smoothness of the effect curve, we may be able to construct more precise estimates. This suggests replacing the $\delta(s)X_{ij}$ term in (8) with some sort of smoothing term. In this paper, we consider the use of natural cubic splines, although other approaches are possible. For an overview of spline-based methods, see Friedman et al.[24] Briefly, given a set of real numbers $k_1 < \ldots < k_d$ called "knots", a cubic spline is a function that is equivalent to a cubic polynomial over any interval $[k_r, k_{r+1}]$ for $r \in \{1, \ldots, d-1\}$. A natural cubic spline is further constrained such that it is continuous, twice continuously differentiable, and linear to the left of the first knot and to the right of the last knot. Through the construction of a so-called spline basis – a set of functions $b_1, \ldots, b_d$ that are applied to the variable of interest – natural cubic splines can be embedded within the linear mixed model framework. Typically, a natural cubic spline with d knots can be represented with d basis functions. In our context, we enforce the additional constraint that the spline must pass through the origin. This leads us to consider the following model, in which the terms $b_1(s), \ldots, b_d(s)$ represent a natural cubic spline basis with d degrees of freedom.

$$g(\mu_{ijk}) = \Gamma_j + [\omega_1 b_1(s) + \ldots + \omega_d b_d(s)]X_{ij} + C_{ijk} \qquad (13)$$

Note that the construction of the spline basis requires the user to specify the number and placement of the knots. A model that uses a basis with $J - 1$ degrees of freedom will yield identical estimation and inference to the ETI model, and so if this model is used, it should be the case that $d < J - 1$. Given that fitting this model yields an estimate of the entire function $s \mapsto \delta(s)$, we could in theory estimate the TATE via integration. However, for simplicity

and to achieve consistency with the other models we consider, we again use a right-hand Riemann sum approximation. This allows us to conduct estimation and inference based on the estimates $\widehat{\omega}_1, \ldots, \widehat{\omega}_d$ of $\omega_1, \ldots, \omega_d$ and the associated covariance matrix estimate. This yields the following estimators, which are analogs of (12):

$$\widehat{\Psi}_{(s_1, s_2]} \equiv \frac{1}{s_2 - s_1} \sum_{r = 1}^{s_2 - s_1} \sum_{k = 1}^{d} \widehat{\omega}_k b_k(s_1 + r)$$

$$\widehat{\Psi}_{s_0} \equiv \sum_{k = 1}^{d} \widehat{\omega}_k b_k(s_0)$$

(14)

The spline basis and the estimators above are easily calculable with existing software, and variance estimation proceeds as before, since these estimators are also linear combinations of the parameter estimates.

Monotone effect curve (MEC) model—It will often be reasonable to assume that the effect curve is monotone; for example, more cardiovascular exercise sustained over a longer period of time will generally lead to more weight loss. When this assumption is plausible, it is natural to wonder whether we can leverage this knowledge to obtain more precise estimates. Constructing models that enforce monotonicity can be done in a number of ways; we introduce one possible model here.

$$g(\mu_{ijk}) = \Gamma_j + \delta\left(\sum_{t = 0}^{s} \alpha_t\right) X_{ij} + C_{ijk}$$

(15)

Above, we let $\alpha_0 \equiv 0$ and constrain $(\alpha_1, \ldots, \alpha_{J-1})$ as a simplex such that $\alpha_t \geq 0$ for $t \in \{1, \ldots, J-1\}$ and $\sum_{t=1}^{J-1} \alpha_t = 1$. This model parameterizes the effect curve as a monotonic step function, but allows for the step function to be either nondecreasing or nonincreasing. This is a constrained nonlinear model, and so more advanced techniques are needed to fit it and estimate the parameters $(\delta, \alpha_1, \ldots, \alpha_{J-1})$. We choose to fit it as a hierarchical Bayesian model using the following prior specification, where $(c_1, \ldots, c_{J-1})$ are fixed constants and $(\delta, \omega, \alpha_1, \ldots, \alpha_{J-1})$ are parameters:

$$\delta \sim \text{Normal}(0, 10^4)$$
$$\omega \sim \text{Uniform}(0.01, 100)$$
$$(\alpha_1, \ldots, \alpha_{J-1}) \sim \text{Dirichlet}(c_1\omega, \ldots, c_{J-1}\omega)$$

(16)

Choosing some values of the constants $(c_1, \ldots, c_{J-1})$ to be larger than the others can be seen as encoding the prior belief that the biggest "jump" in the effect curve will occur in a certain region of the effect curve, or equivalently that the curve will remain relatively flat over a certain region. For example, choosing $c_1 = \ldots = c_{(J-1)/2} = 5$ and $c_{(J-1)/2+1} = \ldots = c_{J-1} = 1$ encodes the prior belief that the biggest jump in the effect curve occurs in the first half of the effect curve (assuming $(J-1)/2$ is an integer); this is the prior we use in this

paper in the simulations and the real data analyses. Choosing $c_1 = \ldots = c_{J-1} = 1$ leads to a symmetric Dirichlet prior, which can be seen as minimally informative in the sense that it doesn't presume in advance that the jump happens at any particular time point.

It is important to note that monotone-constrained function estimators tend to be biased at the endpoints[25]; in our case, the endpoint of the effect curve is precisely the LTE, if it is assumed to exist. The prior described above has the effect of counteracting this bias to an extent by stabilizing the tail of the effect curve. Thus, the results we obtain when using this model are influenced not just by the monotonicity constraint, but also by the informative prior. As we will see in both simulated and real data, different choices of prior may lead to very different estimates, and so we recommend that this model only be used in the context of an exploratory analysis.

In this model, the LTE estimator is simply the posterior mean $\hat{\delta}$. For the TATE, we again use an estimator based on a right-hand Riemann sum. This can be done by first computing the posterior means $\hat{\delta}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{J-1}$, then calculating $\hat{\delta}_1, \ldots, \hat{\delta}_{J-1}$ using the formula $\hat{\delta}_s \equiv \hat{\delta} \sum_{t=1}^{s} \hat{\alpha}_t$, and finally using these to compute the estimators in (12) or (C3). Variance estimation can be done using standard methods for Bayesian inference.

3.3 | Incorporating random effects

In the analysis of stepped wedge trials using an immediate treatment effect model, it is common to include mean-zero random effect terms to capture cluster-level or temporal deviations from the fixed effects. For example, we may include random intercepts to model cluster-level deviations, and we may include random effects at the level of the cluster-by-time interaction (a type of "random time effect") to allow the underlying time trend to differ by cluster. These random effects can be incorporated into models that allow for a time-varying treatment effect in the same way they would be added to a model that assumes an immediate treatment effect. However, more thought is required in terms of incorporating "random treatment effects", which allow for the effect of the treatment to vary between clusters. When the treatment effect does not vary with time, this takes the form of a random coefficient on the treatment indicator variable, indexed by cluster. When the effect of treatment varies by exposure time, there are multiple ways in which random effects could be used to allow the effect curve to differ across clusters. For simplicity, we focus on the case in which the only heterogeneity component other than the random treatment effect is a cluster-level random intercept. The specification given in (17) allows for the "height" of the effect curve to vary between clusters and involves two additional parameters. Essentially, each of the parameters $(\delta_1 \ldots, \delta_{J-1})$ represents the value of the effect curve at a particular exposure time averaged across clusters, and each of the parameters $(\eta_1, \ldots, \eta_I)$ represents the amount by which the entire effect curve is vertically shifted for a given cluster. Other random effects structures that allow for both the "height" and the "shape" of the effect curve to vary between clusters are possible but may be difficult to fit due to their complexity.

$$g(\mu_{ijk}) = \Gamma_j + (\delta_s + \eta_i)X_{ij} + \alpha_i$$

$$\begin{pmatrix} \alpha_1 \\ \eta_1 \end{pmatrix}, \ldots, \begin{pmatrix} \alpha_I \\ \eta_I \end{pmatrix} \overset{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \rho\tau\nu \\ \rho\tau\nu & \nu^2 \end{pmatrix}\right]$$

(17)

The NCS and MEC models can be adapted to include random treatment effects in an analogous manner.

3.4 | Implications for study design and power calculation

In stepped wedge studies, data collection typically stops after all clusters have reached the treatment state, since additional data collection provides little additional information on the treatment effect when an immediate treatment effect model with a saturated study time effect is used for analysis. However, when a model is used that allows for time-varying treatment effects, additional data collection may lead to gains in precision, depending on the estimand of interest. We can gain intuition for why this is the case by considering estimation of the LTE. With a typical stepped wedge design, only one sequence is observed at exposure time $J-1$ (at study time J), and so if the ETI model is used, all the information about the change in the effect curve between times $J-2$ and $J-1$ must come from this sequence alone. However, if we collect additional data at study time step $J+1$, there are now two sequences that have been observed at exposure time $J-1$, and so we should be able to estimate the LTE more precisely, assuming we still think that the long-term effect is reached by exposure time $J-1$. In addition, we could combine the information collected at exposure times $J-1$ and J to improve precision further. Similar logic applies to estimation of the TATE, although we would not expect the gains in precision to be as large. For either estimand, the magnitude of this potential gain in precision can be roughly quantified via simulation (or analytically).

An essential part of stepped wedge trial planning is power (or sample size) calculation, and in the context of time-varying treatment effects, the existing guidance requires modification. While this is not the focus of this paper, we briefly discus two potential approaches to power calculation. The first method is to adapt the existing stepped wedge power formulas, which are based on Wald-type tests. With the IT model, this test involves estimating $\mathrm{Var}(\hat{\delta})$, which is extracted from the estimated covariance matrix of the fixed effects. If the estimand of interest is the PTE or LTE and the ETI model is used, the Wald test will similarly be based on a single model parameter and the test can be constructed in an analogous manner. If the TATE is the estimand of interest and the ETI model is used, the estimate will be a linear combination of the fixed effects, and a variance estimate can be constructed from the assumed covariance matrix, as described in Section 3.2. If a model other than the ETI model is to be used for analysis, the ETI model can still be used for the sake of power calculation, but this may lead to power being underestimated. Note that the variance of the TATE depends only on the study design and the variance components, just as is the case when computing power using the IT model. Furthermore, for a given true TATE value, the variance of the TATE estimate will *not* depend on the underlying effect curve. Therefore, for the sake of power calculation, one only needs to choose a value for the TATE under the

alternative hypothesis and does not need to make additional assumptions about the shape of the effect curve.

The second method is simulation-based power calculation. With this method, the researcher repeatedly simulates the entire experiment and estimates power as the proportion of experiments in which the null hypothesis is rejected. Simulating the experiment will typically involve generating a dataset that mimics the real-world data you expect to collect and running the same analysis you plan on using for the actual trial. Randomness is introduced into the data-generating mechanism by sampling from probability distributions (Normal, Binomial, etc.) or sampling from a fixed population list or sampling frame. While this process requires programming knowledge and can be computationally intensive, it is extremely flexible in that any study design, outcome type, sampling strategy, or analysis technique can be used.

4 | SIMULATION STUDY

We conducted a simulation study designed to assess the performance of the models described in Section 3.2 in a variety of settings. Unless otherwise specified, data were generated according to the following model, a special case of (8):

$$
\begin{aligned}
Y_{ijk} &= \mu + \beta_j + \delta h(s) X_{ij} + \alpha_i + \epsilon_{ijk} \\
(\alpha_1, \ldots, \alpha_I) &\overset{iid}{\sim} N(0, \tau^2) \\
(\epsilon_{111}, \ldots, \epsilon_{IJK}) &\overset{iid}{\sim} N(0, \sigma^2)
\end{aligned}
\tag{18}
$$

The function $s \mapsto h(s)$ represents the effect curve, constrained to start at $(0, 0)$ and achieve a maximum value of 1 on the interval $[0, J - 1]$. For all simulations, we generated data according to four different effect curves, defined as step function approximations to the effect curves a-d in Figure 1. This was done so that when we are comparing TATE estimates to the true values, we eliminate the component of the error due to the step function being an approximation to a smooth curve.

Simulations involved $I = 24$ clusters, $J = 7$ time points, and $K = 20$ individuals per cluster. Parameters were fixed at $\mu = 1$, $\delta = 0.5$, $\sigma = 2$, and $\tau = 0.5$. This results in an ICC of $\rho = 0.059$ and a value of ϕ (as defined in Theorem 1) of $\phi = 0.56$. The study design was balanced and complete, in the sense that there were an equal number of clusters in each sequence and one sequence crossed over at each time point. Data were generated with a linear time trend with $\beta_j = -0.5\left(\frac{j-1}{J-1}\right)$, although the analysis models treat time as categorical.

For the MEC model, we use a Dirichlet$(5\omega, 5\omega, 5\omega, \omega, \omega, \omega)$ prior. As discussed in Section 3.2, this can be seen as encoding the prior belief that the biggest "jump" in the effect curve will occur in the first three exposure time points, or equivalently that it will remain relatively flat after the last three exposure time points.

In all simulations, we considered estimation of $\Psi_{(0,J-1]}$, the TATE between exposure times 0 and $J - 1$, as well as the LTE Ψ_{J-1}, the value of the effect curve at time $J - 1$. Performance

was assessed by estimating bias, 95% confidence interval coverage, and mean squared error (MSE). We also assessed estimation of the entire effect curve by calculating the average pointwise MSE, defined as $\frac{1}{J-1}\sum_{s=1}^{J-1}(\widehat{\Psi}_s - \Psi_s)^2$, and estimated the power of Wald-type hypothesis tests. Each individual statistic was calculated using 1,000 simulation replicates. All simulations were conducted using the R programming language and structured using the *SimEngine* simulation framework.[26]

4.1 | Performance of the IT, ETI, NCS, and MEC models

First, we compare the performance of several of the models described above that are designed to account for a time-varying treatment effect, in addition to the IT model for comparison. Results are given in Figure 3. As expected, the IT model performs well when it is correct; simulation results confirm that the TATE/LTE estimate is unbiased and that nominal 95% coverage is achieved. However when the IT model is not correct, estimates are severely biased and confidence interval coverage is unacceptable. The ETI model will always be correct in the sense that it makes no assumptions about the form of the effect curve. For this reason, it will also typically be the least efficient. We indeed observe that the ETI model leads to unbiased estimates of both the TATE and the LTE; the tiny amount of bias observed is due to Monte Carlo error. The NCS model performs similarly to the ETI model in terms of coverage and MSE, but appears to do slightly worse in terms of bias. The similar performance is not surprising in this case, as there are only six exposure time points in this simulation setup, and so the NCS model involves just two fewer degrees of freedom than the ETI model. The MEC model consistently yields better MSE than the ETI and NCS models for LTE estimation. However, it also leads to highly biased estimates in certain scenarios, and additional simulations (see Appendix B) show that the model is quite sensitive to the choice of prior, in the sense that a minimally-informative prior leads to higher bias and greater undercoverage. Given these drawbacks, we recommend that the MEC model should not be used for confirmatory analyses. Note that "coverage" refers to credible interval coverage in the case of the MEC model.

Figure 4 shows the results of estimating the entire effect curve, assessed via average pointwise MSE, as described above. Overall, the ETI and NCS models perform similarly. The MEC model performs the best for three out of the four effect curves.

Next, using the same set of models as above, we conduct a set of Wald-type hypothesis tests related to the TATE ($H_0 : \Psi_{(0,J-1]} = 0$ vs. $H_1 : \Psi_{(0,J-1]} \neq 0$) and to the LTE ($H_0 : \Psi_{J-1} = 0$ vs. $H_1 : \Psi_{J-1} \neq 0$). Results are given in Figure (5). As expected, tests based on the IT model parameter estimates are the most powerful when the model is correct. When it is incorrect, the power of the IT model suffers dramatically for the lagged (b) and partially convex (d) effect curves. Importantly, this implies that if the IT model is used in the presence of a time-varying effect, there is a serious risk that hypothesis tests will suffer from high type II error rates even if a strong effect is present. The MEC model appears to have the best overall performance, and the ETI and NCS models both perform reasonably well across all scenarios. All tests maintain proper type I error rates when there is no true effect.

In Appendix B, we compare the ETI model to an analogous model that additionally includes a random treatment effect term corresponding to the specification given in (17). We show that the use of a random treatment effect in the analysis model is beneficial in the sense that performance does not suffer when no random treatment effect is present in the data, whereas if it is present in the data, performance improves considerably; this is consistent with previous findings. [20]

4.2 | Effect of adding extra time points

As discussed in section 3.4, when using models that allow for time-varying treatment effects, it may be possible to gain precision by collecting additional data after all clusters have reached the treatment state. Figure 6 shows the results of a set of simulations in which we added either zero, one, or two additional post-treatment time points and analyzed data using the ETI model. The effect curves used to generate the data are the same ones used in all previous simulations, but with $\Psi_{J-1} = \Psi_J = \Psi_{J+1}$; that is, the true effect curves remain flat after exposure time $J - 1$. We continue to define the LTE as Ψ_{J-1}. We observe that adding extra time points enables considerable MSE gains in terms of estimating the LTE, but provides little benefit in terms of estimating the TATE.

5 | DATA ANALYSIS

To illustrate the use of the models described in Section 3.2, we conducted secondary analyses of data from two different stepped wedge trials.

5.1 | Australia weekend services disinvestment trial

The first data analysis is of a stepped wedge trial examining the impact of the disinvestment (removal) of weekend health services from twelve hospital wards in Australia, previously analyzed by Haines et al. (Trial 1).[27] Although the original investigators considered several outcomes, we focus on the the (log) length of hospital stay in days, treated as a continuous variable. This study involved a population of 14,834 individuals, 12 clusters, 6 sequences, and 7 time points.

In our analysis, we fit four different models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with four degrees of freedom (NCS-4), and monotone effect curve (MEC). The treatment of exposure time differs based on which model is used (as described in Section 3.2), whereas study time is modeled as categorical. In this trial, "study time" (the amount of time that has passed since the start of the study) can actually be seen as distinct from "calendar time" due to the complex nature of the study design. For simplicity, and since this distinction is uncommon, we only adjust for study time. The MEC model uses the same prior that was used in simulations. Estimates of the TATE over the entire study ($\Psi_{(0,6]}$) and the LTE (Ψ_6) are given in Figure 7, and estimates of the entire effect curve, along with pointwise confidence bands, are given in Figure 8.

Estimates of the TATE and LTE from the IT model are both much smaller than the corresponding estimates from the three other models. Additionally, the single effect estimate from the IT model is less than each of the point treatment effects estimated from the ETI model. This is an example of the phenomenon described in section 2. The effect

curves estimated from the three models that account for a time-varying treatment effect are all qualitatively similar, which is reassuring. The confidence intervals resulting from the ETI model are the widest, which is as expected since this model makes no assumptions about the shape of the effect curve. The NCS model performs almost identically to the ETI model, which is not surprising for reasons discussed in section 4.1. The MEC model performs somewhat similarly to the ETI model, but with narrower confidence intervals. As a sensitivity analysis, we also ran the MEC model with a (minimally informative) symmetric Dirichlet prior (i.e. $c_1 = \ldots = c_{j-1} = 1$). This resulted in a similar TATE estimate of 0.18 (95% CI: 0.09–0.28) but a much higher LTE estimate of 0.30 (95%CI: 0.15–0.45), reinforcing the knowledge that the MEC model can be quite sensitive to prior selection.

5.2 | Washington State expedited partner treatment trial

Next, we conducted a secondary analysis of data from the Washington State Community-Level Expedited Partner Treatment (EPT) Randomized Trial.[28] This trial sought to test the effect of EPT, an intervention in which the sex partners of individuals with sexually transmitted infections are treated without medical evaluation, on rates of chlamydia and gonorrhea. This study involved a population of 390,675 individuals, 22 clusters, 4 sequences, and 4 time points, although measurements were taken more frequently (at 15 time points). In our analysis, we applied the same set of models used for the first data analysis, but using a binomial GLM with a logit link and random intercepts for both cluster and site. Again, time was modeled as categorical. Estimates of the TATE over the entire study ($\Psi_{(0,14]}$) and the LTE (Ψ_{14}) are given in Figure 9, and estimates of the entire effect curve, along with pointwise confidence bands, are given in Figure 10.

In this analysis, we again observe that the effect estimated from the IT model is less than each of the point treatment effects estimated from the ETI model. Furthermore, the trend is estimated in the opposite direction (odds ratio < 1) in the IT model relative to the models that account for a time-varying treatment effect (each odds ratio > 1). The IT estimate of the TATE is less than one, suggesting that the intervention has a small (nonsignificant) benefit, but the TATE estimates from the other models are all greater than one. However, all of the confidence intervals are quite wide and contain one (the null value), and so none of the associated Wald-type tests reject the null hypothesis of no effect, for both the TATE and the LTE.

In this analysis, we see the smoothing behavior of the NCS model in action when comparing its estimated effect curve to that of the ETI model. As described earlier, this is expected, since with 14 time points there is a much greater difference in terms of the degrees of freedom between the ETI and NCS models. While one might expect the MEC model to yield an estimated effect curve similar to that of the NCS model, we do not observe this here; this is due to the effects of the prior. When we instead fit the MEC model with a symmetric Dirichlet prior (as we did as a sensitivity analysis for the previous dataset), we obtain a TATE estimate of 1.17 (95% CI: 098–1.41) and an LTE estimate of 1.36 (95%CI: 0.97–1.91), both of which are quite different than the results in Figure 9. These TATE and LTE estimates, as well as the estimate of the entire effect curve, are similar to those of the NCS model. Again, this demonstrates the sensitivity of the MEC model to the priors.

6 | DISCUSSION

In this paper, we characterize the risks of not accounting for a time-varying treatment effect in the context of a stepped wedge design, introduce terminology for describing estimands of interest, introduce several models that can be used for estimation and hypothesis testing, evaluate the operating characteristics of these models via simulation, and demonstrate their use on two real datasets.

6.1 | The immediate treatment effect model and estimands of interest

With the immediate treatment effect model, the researcher must assume that the full effect of the treatment is reached immediately, as if the effect can be instantly turned on by flipping a light switch. However, if a time-varying treatment effect is present and the IT model is used, estimates of both the TATE and the LTE can be severely biased, confidence interval coverage can be unacceptably low, and MSE can be very high. As noted in section 2, the expectation of the IT treatment effect estimator can be represented as a weighted sum of the individual point treatment effects $\delta(1)$, ..., $\delta(J-1)$, but with some negative weights. This implies that under certain effect curves, TATE and LTE estimates resulting from the use of the IT model can converge to a value of the opposite sign as the true parameters. This result is quite surprising, and although similar phenomena have been examined in the context of two-way fixed effects designs[29,30], this has never previously been discussed in the stepped wedge literature. Because the impact of this form of model misspecification on estimation and inference is so severe, we recommend that the IT model should not be used moving forward for confirmatory analyses unless the assumption of an immediate treatment effect is justifiable based on contextual knowledge of the intervention. Even when the assumption is justifiable, a sensitivity analysis should be conducted using a model that allows for a time-varying treatment effect. Furthermore, it may be worth reanalyzing data from past stepped wedge trials in which the immediate treatment effect assumption is questionable.

It is natural to wonder why the counterintuitive behavior of the IT model occurs. To gain some intuition, we note that in a stepped wedge trial with Q sequences, all Q will be observed at exposure time 1, but only $Q-1$ will be observed at exposure time 2, only $Q-2$ will be observed at exposure time 3, and so on. This means that as exposure time increases, the amount of information available in the data decreases, and intuitively this is why the weights given in equation (5) are decreasing in s, as well as why the weight corresponding to exposure time 1 is always the largest. This is consistent with previous work that has identified that observations with exposure time 1 contribute the most to estimation of the treatment effect in IT models.[31,32] The reason why some weights can be negative is less intuitive, but insight can be gained by drawing parallels to results around optimal linear combinations of correlated unbiased estimators. When estimating a parameter μ using an estimator of the form $\tilde{\mu} \equiv \alpha\hat{\mu}_1 + (1-\alpha)\hat{\mu}_2$, where $\hat{\mu}_1$ and $\hat{\mu}_2$ are both unbiased for μ, Samuel-Cahn showed that the optimal value of α can be expressed in terms of the correlation $\mathrm{Cor}(\hat{\mu}_1,\hat{\mu}_2)$ and the ratio $\mathrm{Var}(\hat{\mu}_1)/\mathrm{Var}(\hat{\mu}_2)$, and that the optimal choice of α (in terms of minimizing $\mathrm{Var}(\tilde{\mu})$) may be greater than one or less than zero.[33] For example, when estimating the location parameter of a Uniform(0,1) distribution using the sample mean $\hat{\mu}_1$ and the sample median $\hat{\mu}_2$, the optimal combination is $1.5\hat{\mu}_1 - 0.5\hat{\mu}_2$. If we use

the ETI model to analyze a simple stepped wedge design involving two sequences and three time periods, the optimal choice of α for using an estimator of the form $\alpha\hat{\delta}_1 + (1 - \alpha)\hat{\delta}_2$ to estimate the true treatment effect δ (assuming the IT model is correct), as calculated from the Samuel-Cahn result, turns out to be identical to the weight $w(2, \phi, 1)$ obtained from equation (5). This suggests that the IT model is implicitly leveraging the assumption of an immediate and constant treatment effect to obtain an optimal estimator.

In the case of an immediate treatment effect, the effect curve is constant over time, and so the phrase "treatment effect" is well-defined. However, when the effect curve changes over time, this phrase is ambiguous and researchers must take care to specify precisely what they are estimating. We defined the time-averaged treatment effect (TATE), the point treatment effect (PTE), and the long-term treatment effect (LTE) to distinguish between the different estimands a researcher may be interested in. We expect that the LTE will be of interest in settings in which it is assumed to exist, as many interventions seek to cause long-lasting change. However, when the assumption that a long-term effect exists (and is reached by the end of the study) cannot be made, the TATE and the PTE are appealing estimands. We view the PTE as relevant if there is a specific time point of scientific interest, whereas the TATE is relevant if this is not the case. Additionally, the TATE is an appropriate estimand if interest lies in the average effect after a prespecified washout or transition period has taken place. The TATE also likely reflects what researchers previously thought they were estimating when using an IT model in a scenario with a time-varying treatment effect. Furthermore, for a given study design and effect size, it will generally be the case that the power to reject the null hypothesis related to the TATE (over the exposure period or after a short washout period) will be greater than the corresponding test related to a particular PTE or the LTE, and so constraints related to funding or logistics may dictate what is testable and what is not. However, as discussed in section 3.2, if a researcher is willing to assume that the long-term effect is reached by some time point $J - m$ (with $m > 1$) before the end of the study, the TATE estimator $\widehat{\Psi}_{(J - m, J - 1]}$ can be used to estimate the LTE Ψ_{J-1} to gain precision.

When considering the TATE, an alternative definition we considered instead of (9) was $(s_2 - s_1)^{-1} \sum_{s = s_1 + 1}^{s_2} \delta(s)$, the average of the true point treatment effects at the study measurement times. We chose to not use this definition for two main reasons. First, doing so would imply that the estimand of interest is entangled with the study design, in the sense that the quantity and meaning of the $\delta(s)$ terms directly depends on the number of time steps in the study and the spacing between the steps. Second, since time is continuous in reality, we believe that an integral-based average is conceptually closer to what researchers will typically be interested in. However, the estimands may depend on the study design in practice, in the sense that possible choices for $(s_1, s_2]$ in $\Psi_{(s_1, s_2]}$ or s_0 in Ψ_{s_0} are limited by the number and length of time steps in the study design. Similarly, the effect curve flattening assumption necessary for the LTE to be well-defined requires the study to be of a sufficient length to achieve flattening.

6.2 | Features unique to the setting of time-varying treatment effects

For the models that account for a time-varying treatment effect, it will generally be the case that the variance of the PTE increases with exposure time, since the available information decreases as exposure time increases, as described above. We also note that some of the existing guidance for the design of stepped wedge trials requires modification when there is a time-varying treatment effect. Normally, data collection stops after all clusters are observed in the treatment state. However, in our simulations, we explored the potential for collecting data at additional time points and showed that this can lead to increased precision, particularly if the LTE is of interest. Additionally, care should be taken when considering designs involving different time step lengths. For example, consider a design in which all clusters are measured at baseline, at three months, and at eight months (in terms of study time). In this case, the first three measurements for sequence 1 will occur at exposure times 0, 3, and 8, whereas the first three measurements for sequence 2 will occur at exposure times 0, 0, and 5. This is not an issue for the IT model, but an ETI model will require more exposure time parameters (i.e. the δ_S terms), since each of these parameters represents the treatment effect at a specific exposure time. The NCS model is very appealing in this scenario, since it allows the researcher to explicitly limit the number of model parameters; that is, the number of spline basis functions can remain constant even with a growing number of unique exposure time values.

6.3 | Which model to choose

The ETI model is saturated with respect to exposure time, and thus makes no assumptions about the shape of the effect curve (assuming measurements are taken at regular intervals); in this sense, it is the most robust model that we consider. If the study time trend is correctly specified and a linear link function is used, the ETI model will always yield unbiased estimates of the PTE and the LTE, as well as (aside from the error due to the use of a Riemann sum in approximating the effect curve) the TATE for any time interval. If instead a nonlinear link function is used, estimates will be unbiased if the heterogeneity model component is also correctly specified. This is true regardless of the properties of the design (number of clusters, number of time points, etc.). Additionally, if interest lies in estimation of the TATE after a washout or transition period, the ETI model is preferable to using a model that assumes a washout period followed by an immediate treatment effect, as such a model will suffer from issues similar to those discussed in section 2, if the assumptions are incorrect. However, the ETI model will also lead to estimators with the greatest variance. It will sometimes be impractical to estimate the LTE using this model, since the information about the change in the effect curve between the second-to-last and last time points comes from just a single sequence. This is a problem not only in terms of variance, but also in terms of generalizability, especially if there are only a few clusters per sequence. This being said, we see that the LTE estimate in the Australia disinvestment trial is statistically significant for the ETI model (as well as for the other models considered). Overall, we anticipate that the ETI model will be a practical choice in many applications.

With the NCS model, although we observe the smoothing effect in the WA State data analysis, we don't see evidence of improved estimation or inference in our simulation study. However, we still view this model as useful, particularly when the entire effect curve is of

interest. Further research is needed to determine optimal knot number and placement. Also, as mentioned above, the NCS model is well-suited to trials in which the lengths of the time steps differ, since otherwise the number of exposure time parameters (i.e. the δ_s terms) could be very high, yielding TATE estimators with high variance.

We showed that the use of a random treatment effect is beneficial in the ETI model, but the same conclusion applies to the NCS and MEC models. There is little penalty for including a random treatment effect if it is feasible to do so, since performance does not suffer when no random treatment effect is present in the data, but if it is present in the data we see a considerable performance improvement, particularly in terms of coverage.

Next, we discuss the MEC model. While this model performs well in certain simulation scenarios, it is highly sensitive to the choice of prior, as we demonstrated in both simulations and real data. As mentioned in section 3.2, estimates obtained by this model are influenced not just by the monotonicity constraint, but also by the informative prior. In general, since this model can lead to high bias and since the choice of prior has a considerable influence on the resulting estimates, we recommend that this model is only used for exploratory analyses.

In summary, if the TATE is the estimand of interest, then the ETI model is a good all-purpose choice, as its properties are well-understood and it minimizes the need to make assumptions about the shape of the effect curve. The NCS model is a good alternative if there is a large number of time points or unevenly-spaced steps in the design. If the LTE is the estimand of interest, the ETI and NCS models are again both good choices, but the researcher should carefully consider whether a large enough sample size is possible and should also consider collecting additional post-treatment data to gain precision. For both estimands, the IT model should only be used if the researcher is highly confident that the treatment effect does not vary with time, and if it is used, the ETI model should be fit as a sensitivity analysis.

6.4 | Assumptions and limitations

A critical assumption made throughout this work is that the treatment effect varies as a function of exposure time, but *not* as a function of study time. While this assumption will hold in many scientific settings, researchers should carefully consider whether it holds in each particular study. A notable example of when this assumption would fail is when there is a major external shock (e.g. a natural disaster) that occurs in the middle of the study period that affects implementation of the intervention. Also, while all of the analysis models we consider use random effects to account for the correlation structure of the data, other approaches are possible.[34,35,36,37] Most of these alternative approaches, such as GEE, differ mainly in terms of how the dependence between observations is handled. Thus, by making analogous modifications to the fixed effects structures, these approaches can be adapted to settings in which there is a time-varying treatment effect. However, we did not study these approaches in depth. Another limitation of this work is that the set of simulation scenarios we considered was limited, and represents a small fraction of the set of possible data-generating mechanisms, in terms of true effect curves, ICC values, outcome variable types, covariance structures, etc.

6.5 | Future research directions

We see a number of open questions that could be explored through future methodological work. First, as mentioned in section 3.4, the use of a time-varying treatment effect model will impact statistical power, and more research is needed to properly characterize this impact. Second, the setting of a time-varying treatment effect may lead to additional implications on study design other than the possibility of adding additional time points. These implications will likely differ depending on the estimand of interest. The topic of model misspecification is also worthy of further exploration. In section 2, we explored the behavior of the IT model under a time-varying treatment effect but focused on the Hussey and Hughes model; it would be informative to perform similar analyses for other other models, correlation structures, and sampling schemes (e.g. cohort sampling instead of cross-sectional sampling). It would similarly be useful to examine how the models we study here perform when the treatment effect varies as a function of study time rather than exposure time, as well as to construct models that handle this scenario. Finally, it would be worth doing further research into alternative ways to estimate the time trend in the context of a time-varying treatment effect. Since moving from the IT model to the ETI model results in estimators with higher variance, imposing a more restrictive model for the time trend could potentially help to gain back some precision. For example, it would be straightforward to use a natural cubic spline (or other smoothing estimator) to estimate the time trend. Other approaches that have been considered include the use of linear[17] and quadratic[12] time trends, or the exclusion of time trends entirely for short trials.[38] It would be useful to understand what precision gains are possible, as well as the costs associated with time trend misspecification.

7 | CONCLUSIONS

If a stepped wedge trial is testing a treatment that varies as a function of exposure time, the use of a model that assumes an immediate treatment effect can lead to serious errors in both estimation and inference. We introduced several models that account for time-varying treatment effects and evaluated their operating characteristics via simulation. We recommend that a model that accounts for time-varying treatment effects is always used in stepped wedge trials moving forward unless the researcher has compelling evidence that the treatment effect is immediate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX

A   BEHAVIOR OF THE IT MODEL UNDER ALTERNATIVE CORRELATION STRUCTURES

In section 2, we investigated the behavior of the treatment effect estimator resulting from the IT model (specifically the Hussey and Hughes model), and showed that the expectation of this estimator can be expressed as a weighted sum of the individual point treatment effects $\{\delta(1), \ldots, \delta(J-1)\}$. In this section, we provide additional results illustrating the behavior of these weights for several correlation structures. This analysis shows that the behavior of the IT model treatment effect estimator observed in section 2 is not restricted to the model in which heterogeneity is modeled as a single cluster-level random intercept.

A.1   Exchangeable correlation structure

First, we use equation (5) to show how the weights vary as a function of ϕ for select values of Q under the assumption of an exchangeable correlation matrix, the structure implied by model (1). Results are given in Figure A1.
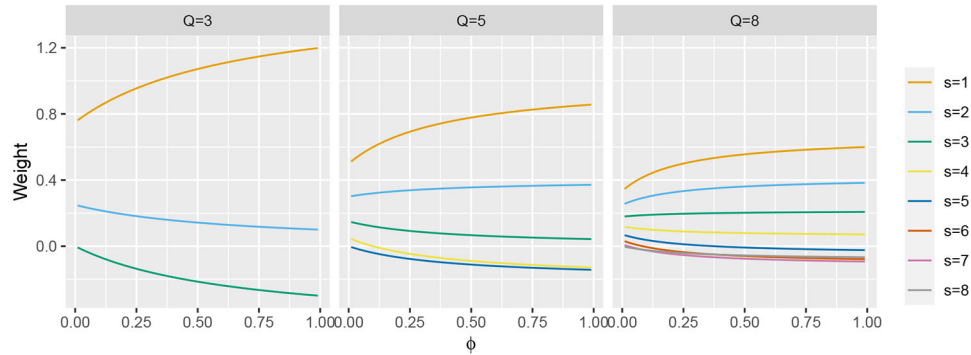

FIGURE A1.
Weights $w(Q, \phi, s)$ plotted as a function of ϕ for $Q \in \{3, 5, 8\}$.

We observe that the weight $w(Q, \phi, 1)$, which corresponds to the point treatment effect after a single time step, receives the most weight, regardless of the value of Q or ϕ. Additionally, in the $Q = 8$ case, we see that half of the weights are negative as $\phi \to 1$. This implies that, under certain true effect curves, the expected value of the IT model treatment effect estimator may be of the opposite sign as the true TATE or LTE. As noted in section 2, the value of ϕ can be quite large if the number of individuals per cluster is large, even if ρ is small, and so we expect a broad range of ϕ values to be encountered in real trials.

A.2 Correlation structure implied by a random time effect

Next, we produce analogs of Figure A1 for several alternative correlation structures. Rather than deriving analytical results, we use Mathematica[39] software to calculate treatment effects estimates as a function of $\{\delta(1), \ldots, \delta(J-1)\}$ by computing the weighted least squares estimator. We consider a design involving three sequences with one cluster each and four time points.

We start with the correlation structure implied by a random time effect. This model is the same as the Hussey and Hughes model, but with the addition of a Normal random effect indexed by both cluster and time representing a unique cluster-by-time interaction. The resulting correlation matrix involves two parameters; ρ_w represents the correlation between two observations within a cluster measured at the same time point and ρ_b represents the correlation between two observations within a cluster measured at different time points. This structure is depicted in (A1), adapted from Li et al.[20], for a trial involving three time points and two individuals per time point.

$$\begin{pmatrix} 1 & \rho_w & \rho_b & \rho_b & \rho_b & \rho_b \\ \rho_w & 1 & \rho_b & \rho_b & \rho_b & \rho_b \\ \rho_b & \rho_b & 1 & \rho_w & \rho_b & \rho_b \\ \rho_b & \rho_b & \rho_w & 1 & \rho_b & \rho_b \\ \rho_b & \rho_b & \rho_b & \rho_b & 1 & \rho_w \\ \rho_b & \rho_b & \rho_b & \rho_b & \rho_w & 1 \end{pmatrix} \tag{A1}$$

The resulting weights are depicting in Figure A2 for a design involving two individuals per cluster, for select values of ρ_w and ρ_b.
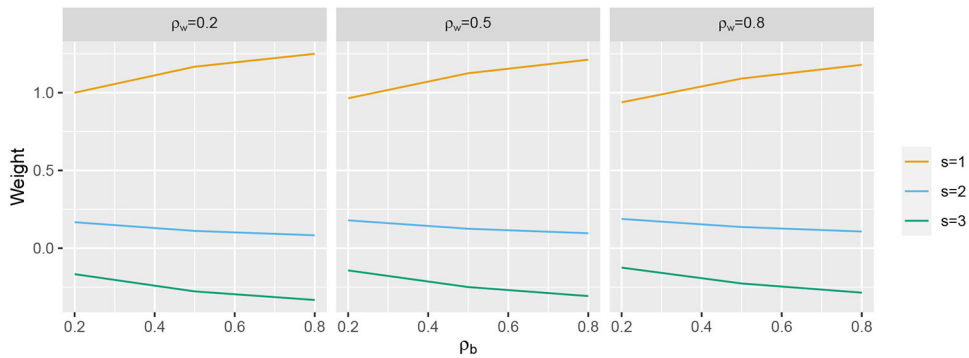

FIGURE A2.
Weights plotted for select values of ρ_w and ρ_b for $Q = 3$.

A.3 Correlation structure implied by a random treatment effect

Next, we computed the weights for the correlation structure implied by a random treatment effect (we discussed random treatment effects in section 3.3). The resulting correlation matrix involves three parameters; ρ_0 represents the correlation between two observations in the control state, ρ_1 represents the correlation between two observations in the treatment

state, and ρ_{10} represents the correlation between two observations, one of which is in the control state and one of which is in the treatment state. The resulting weights are depicting in Figure A3 for select values of ρ_0, ρ_1, and ρ_{10}.
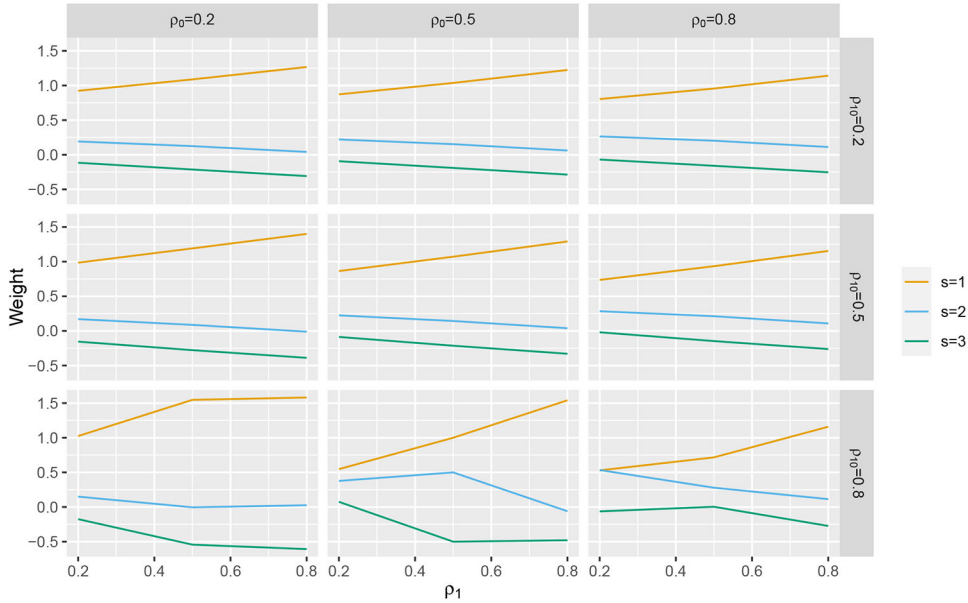

FIGURE A3.
Weights plotted for select values of ρ_0, ρ_1, and ρ_{10} for $Q = 3$.

We see that a similar pattern of weights holds for all three correlation structures considered. This indicates that the problem of IT model misspecification discussed in section 2 is not restricted to the Hussey and Hughes model, but is a general problem with models that assume an immediate treatment effect.

B ADDITIONAL SIMULATION RESULTS

B.1 Performance of models in the presence of random treatment effects

In the set of simulations summarized in Figure B4, data were generated according to the same model used for the simulations in section 4.1, except additionally with a cluster-level random treatment effect generated according to (17) with $\nu = 1$ and $\rho = -0.2$. Data were analyzed using an ETI model that does not account for random treatment effects, and a second ETI model that implements the random treatment effect structure given in (17); the latter is abbreviated as "ETI-RTE". The results are consistent with findings from previous studies on variance structures in the context of stepped wedge designs.[40,41] When a random treatment effect is present in the data but we fail to account for it, we suffer considerably in terms of coverage and MSE.
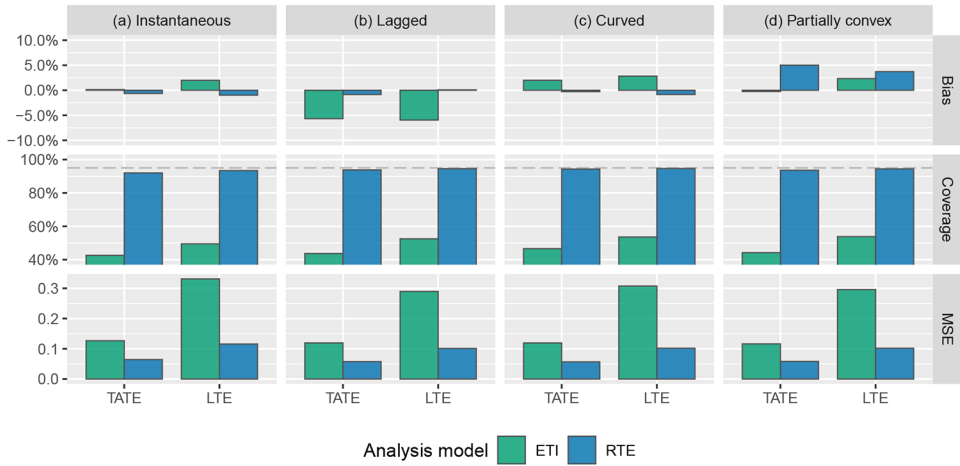
FIGURE B4.

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,J-1)}$) and LTE (Ψ_{J-1}) when data are generated with random treatment effects, using the exposure time indicator model (ETI) and the exposure time indicator model with a random treatment effect (ETI-RTE)

We also performed an analogous set of simulations in which data were instead generated without a random treatment effect (i.e. with $\nu = 0$). In these simulations, the ETI and ETI-RTE models performed similarly across all three metrics; results are shown in Figure B5.

Together, these results suggest that random treatment effect terms should be used whenever possible, since very little is lost in terms of inferential performance when no random treatment effect is present in the data, whereas if it is present in the data we achieve much better performance.

B.2 Use of a symmetric Dirichlet prior for the MEC model

As noted in the discussion, if we use a symmetric Dirichlet prior (i.e. Dirichlet(ω, ω, ω, ω, ω, ω)) in simulations instead of a Dirichlet(5ω, 5ω, 5ω, ω, ω, ω) prior, we see much different behavior of the MEC model. Estimates are more biased and undercoverage is generally worse. We also don't see gains in MSE that are as dramatic, and in some cases, MSE is worse in the MEC model than it is in the ETI model. Results are given in Figure B6.
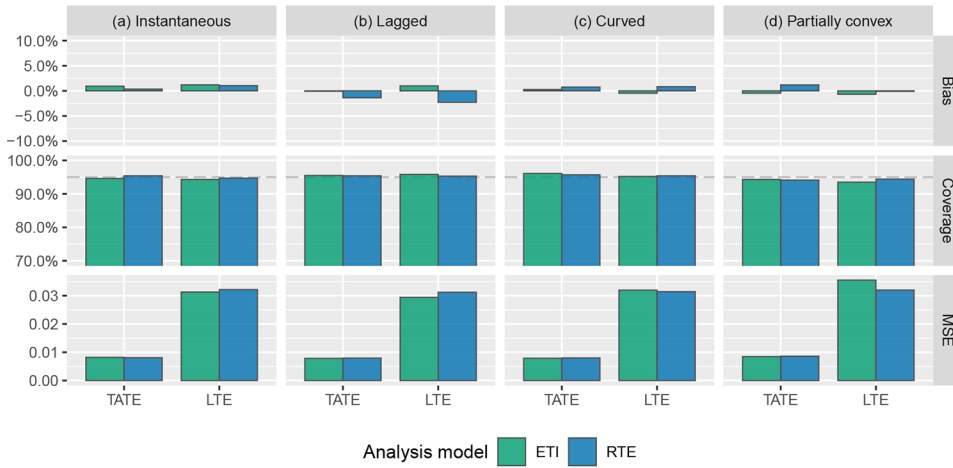
FIGURE B5.

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,J-1]}$) and LTE (Ψ_{J-1}) when data are generated without random treatment effects, using the exposure time indicator model (ETI) and the exposure time indicator model with a random treatment effect (ETI-RTE)
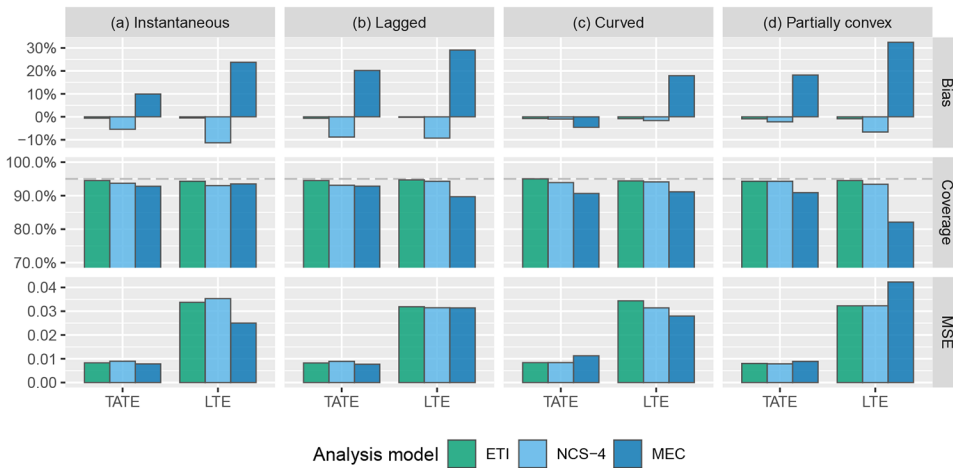

FIGURE B6.

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,J-1]}$) and LTE (Ψ_{J-1}) using the following four models: exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve with a symmetric Dirichlet prior (MEC)

C USE OF TRAPEZOIDAL TATE ESTIMATORS

As noted in section 3.2, the time-averaged treatment effect (TATE) is an integral which must be approximated using measurements taken in discrete time. Throughout the paper, we chose to use TATE estimators based on right-hand Riemann sums, due to their simplicity and to facilitate more natural performance comparisons between different effect curves.

However, it is also worth considering estimators based on trapezoidal Riemann sums. For the ETI model, the right-hand Riemann sum estimator was given by the following:

$$\widehat{\Psi}_{(s_1, s_2]} \equiv \frac{1}{s_2 - s_1} \sum_{r=1}^{s_2 - s_1} \hat{\delta}_{s_1 + r} \tag{C2}$$

The corresponding estimator based on a trapezoidal Riemann sum is given in (C3) below, where for notational convenience we set $\hat{\delta}_0 \equiv 0$:

$$\widehat{\Psi}^*_{(s_1, s_2]} \equiv \frac{1}{s_2 - s_1} \sum_{r=1}^{s_2 - s_1} \frac{\hat{\delta}_{s_1 + r - 1} + \hat{\delta}_{s_1 + r}}{2} \tag{C3}$$

It is informative to compare the variances of these two estimators for the case in which $s_1 = 0$. We can write the following:

$$\begin{aligned}
\text{Var}\left(\widehat{\Psi}_{[0, s_2]}\right) &= \frac{1}{s_2^2} \text{Var}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r + \hat{\delta}_{s_2} \right) = \frac{1}{s_2^2} \\
&\left[\text{Var}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r \right) + \text{Var}\left(\hat{\delta}_{s_2} \right) + 2 \, \text{Cov}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r, \hat{\delta}_{s_2} \right) \right] \\
\text{Var}\left(\widehat{\Psi}^*_{[0, s_2]}\right) &= \frac{1}{s_2^2} \text{Var}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r + \frac{1}{2} \hat{\delta}_{s_2} \right) = \frac{1}{s_2^2} \\
&\left[\text{Var}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r \right) + \frac{1}{4} \text{Var}\left(\hat{\delta}_{s_2} \right) + \text{Cov}\left(\sum_{r=1}^{s_2 - 1} \hat{\delta}_r, \hat{\delta}_{s_2} \right) \right]
\end{aligned} \tag{C4}$$

Simulation results suggest that the covariance term in (C4) is often positive, and so it is reasonable to expect the variance of the trapezoidal estimator to be greater than the variance of the right-hand Riemann estimator. This intuitively makes sense, since the two differ only by a factor of $\frac{1}{2}\hat{\delta}_{s_2}$, and $\hat{\delta}_{s_2}$ will have higher variance than any other $\hat{\delta}_t$ term for $t < s_2$.

Therefore, it is worth considering use of the trapezoidal estimators if it is reasonable to expect the effect curve to be somewhat smooth. Although we focused here on the ETI model, similar trapezoidal estimators can be formed for the other models we considered in this paper.

D PROOFS

Suppose we have a standard stepped wedge design with Q sequences and $J = Q + 1$ time points. Consider the mixed model specified in (10) with $\Gamma_j = \mu + \beta_j$ and $C_{ijk} = \alpha_i$, where $\alpha_1, \ldots, \alpha_I \overset{iid}{\sim} N(0, \tau^2)$. Also assume that, conditional on random effects, individual outcomes are independent and normally distributed with variance σ^2. For now, assume that σ and τ are known, and define $\phi \equiv \tau^2/(\tau^2 + \sigma^2/K)$. From standard results in mixed model theory, we

know that the weighted least squares estimator of the fixed effects $(\mu, \beta_2, ..., \beta_J, \delta)$ is given by the following, where X_q is the design matrix for an individual assigned to sequence q, V is the covariance matrix for a single cluster, and $\bar{Y}_q \equiv (\bar{Y}_q 1, ..., \bar{Y}_q J)'$, with $\bar{Y}_q j$ representing the mean responses among individuals in sequence q at time j:

$$(\hat{\mu}, \hat{\beta}_2, ..., \hat{\beta}_J, \hat{\delta})' = \left(\sum_{q=1}^{Q} X_q' V^{-1} X_q \right)^{-1} \sum_{q=1}^{Q} X_q' V^{-1} \bar{Y}_q$$

Using the form of the covariance matrix $V \equiv \sigma^2 ((1 - \phi) I_J + \phi \mathbb{1}_J)$ implied by the chosen mixed model, where I_J represents a $J \times J$ identity matrix and $\mathbb{1}_J$ represents a $J \times J$ matrix of all ones, calculations done in Matthews and Forbes[18] show that the treatment effect estimator $\hat{\delta}$ can be expressed as follows:

$$\hat{\delta} = \frac{Q \sum_{q=1}^{Q} D_{1q}' V^{-1} Y_q - C_1' V^{-1} \bar{Y}_C}{Q \sum_{q=1}^{Q} D_{1q}' V^{-1} D_{1q} - C_1' V^{-1} C_1}$$

Above, $\bar{Y}_C \equiv (\sum_{q=1}^{Q} \bar{Y}_q 1, ..., \sum_{q=1}^{Q} \bar{Y}_q J)'$, D_{1q} is the treatment allocation vector for sequence q, and C_1 is the vector $(0, 1, ..., Q)'$.

We proceed by analyzing each of the four terms $\sum_{q=1}^{Q} D_{1q}' V^{-1} Y_q$, $C_1' V^{-1} \bar{Y}_C$, $\sum_{q=1}^{Q} D_{1q}' V^{-1} D_{1q}$, and $C_1' V^{-1} C_1$ separately.

We start with the first equality and note the following, where we use the fact that $V^{-1} = \frac{1}{1 - \phi} I_J - \frac{\phi}{(1 - \phi)(1 + J\phi - \phi)} \mathbb{1}_J$ and where $(0, ..., 0, 1, ..., 1)$ represents a vector of q zeros followed by $J - q$ ones:

$$D_{1q}' V^{-1} \bar{Y}_q = \left[\frac{1}{1 - \phi} (0, ..., 0, 1, ..., 1) - \frac{\phi(Q - q + 1)}{(1 - \phi)(1 + J\phi - \phi)} (1, ..., 1, 1, ..., 1) \right] (\bar{Y}_q 1, ..., \bar{Y}_q J)'$$

$$= \frac{1}{1 - \phi} \sum_{j=1}^{J} \left[I\{j > q\} - \frac{\phi(Q - q + 1)}{1 + \phi Q} \right] \bar{Y}_q j$$

Above (and throughout), we use the simple fact that $Q + 1 = J$. The equation above allows us to write the following:

$$\sum_{q=1}^{Q} D_{1q}' V^{-1} \bar{Y}_q = \frac{1}{1 - \phi} \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[I\{j > q\} - \frac{\phi(Q - q + 1)}{1 + \phi Q} \right] \bar{Y}_q j$$

Next, we consider the second term:

$$C_1' V^{-1} \bar{Y}_C = (0, 1, ..., Q) \left(\frac{1}{1 - \phi} I_J - \frac{\phi}{(1 - \phi)(1 + J\phi - \phi)} \mathbb{1}_J \right) \bar{Y}_C$$

$$= \left[\frac{1}{1 - \phi} (0, 1, ..., Q) - \frac{\phi Q(Q + 1)}{2(1 - \phi)(1 + J\phi - \phi)} (1, ..., 1) \right] \bar{Y}_C$$

$$= \frac{1}{1 - \phi} \left(0 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)}, 1 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)}, ..., Q - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)} \right) \left(\sum_{q = 1}^{Q} \bar{Y}_{q1}, ..., \sum_{q = 1}^{Q} \bar{Y}_{qJ} \right)'$$

$$= \frac{1}{1 - \phi} \sum_{q = 1}^{Q} \sum_{j = 1}^{J} \left[j - 1 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)} \right] \bar{Y}_{qj}$$

We proceed with the third equality and note the following:

$$D_{1q} V^{-1} D_{1q} = \left[\frac{1}{1 - \phi} (0, ..., 0, 1, ..., 1) - \frac{\phi(Q - q + 1)}{(1 - \phi)(1 + J\phi - \phi)} (1, ..., 1, 1, ..., 1) \right] (0, ..., 0, 1, ..., 1)'$$

$$= \frac{J - q}{1 - \phi} \left[1 - \frac{\phi(Q - q + 1)}{1 + J\phi - \phi} \right]$$

$$= \frac{(Q - q + 1)(1 + q\phi - \phi)}{(1 - \phi)(1 + Q\phi)}$$

This allows us to express the sum as follows, where we use the identities $\sum_{i = 1}^{n} i = \frac{n(n + 1)}{2}$ and $\sum_{i = 1}^{n} i^2 = \frac{n(n + 1)(2n + 1)}{6}$:

$$\sum_{q = 1}^{Q} D_{1q} V^{-1} D_{1q} = \frac{1}{(1 - \phi)(1 + Q\phi)} \sum_{q = 1}^{Q} (Q - q + 1)(1 + q\phi - \phi)$$

$$= \frac{1}{(1 - \phi)(1 + Q\phi)} \left[\sum_{q = 1}^{Q} (1 - \phi(1 + Q) + Q) + ((Q + 2)\phi - 1) \sum_{q = 1}^{Q} q - \phi \sum_{q = 1}^{Q} q^2 \right]$$

$$= \frac{1}{(1 - \phi)(1 + Q\phi)} \left[Q(1 - \phi(1 + Q) + Q) + ((Q + 2)\phi - 1) \frac{Q(Q + 1)}{2} - \phi \frac{Q(Q + 1)(2Q + 1)}{6} \right]$$

$$= \frac{Q(Q + 1)}{(1 - \phi)(1 + Q\phi)} \left[\frac{1}{2} + \frac{\phi(Q - 1)}{6} \right]$$

Finally, the fourth term can be expressed as follows, where we again use the summation identities that we used to analyze the third term:

$$C_1' V^{-1} C_1 = \frac{1}{1 - \phi} \left(0 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)}, 1 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)}, ..., Q - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)} \right) (0, 1, ..., Q)'$$

$$= \frac{1}{1 - \phi} \sum_{q = 0}^{Q} q \left[q - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)} \right]$$

$$= \frac{1}{1 - \phi} \left[\sum_{q = 0}^{Q} q^2 - \frac{\phi Q(Q + 1)}{2(1 + Q\phi)} \sum_{q = 0}^{Q} q \right]$$

$$= \frac{Q(Q + 1)}{1 - \phi} \left[\frac{\phi Q(Q - 1) + 4Q + 2}{12(1 + Q\phi)} \right]$$

Putting these four equalities together, we have the following result:

$$\hat{\delta} = \frac{12(1 + \phi Q)}{Q(Q + 1)(\phi Q^2 + 2Q - \phi Q - 2)} \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] \bar{Y}_{qj}$$

This is the first result of our theorem. To prove the second result, we start by writing the following:

$$E[\hat{\delta}] = \frac{12(1 + \phi Q)}{Q(Q + 1)(\phi Q^2 + 2Q - \phi Q - 2)} \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] E[\bar{Y}_{qj}]$$

Plugging in $E[\bar{Y}_{qj}] = \mu + \beta_j + I\{j > q\}\delta_{j-q}$ (with $\beta_1 \equiv 0$) and setting $M \equiv \dfrac{12(1 + \phi Q)}{Q(Q + 1)(\phi Q^2 + 2Q - \phi Q - 2)}$, the above is equal to:

$$M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] (\mu + \beta_j + I\{j > q\}\delta_{j-q})$$

$$= M\mu \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] +$$

$$M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] \beta_j +$$

$$M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] I\{j > q\}\delta_{j-q}$$

Next, we show that each of the first two terms is equal to zero. We start with the first term and write:

$$M\mu \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right]$$

$$= M\mu \left[\sum_{q=1}^{Q} \sum_{j=1}^{J} QI\{j > q\} - \sum_{q=1}^{Q} \sum_{j=1}^{J} j + QJ + \sum_{q=1}^{Q} \sum_{j=1}^{J} \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right]$$

$$= M\mu \left[\frac{Q^2(Q + 1)}{2} - \frac{Q(Q + 1)(Q + 2)}{2} + Q(Q + 1) + \frac{\phi Q(Q + 1)}{2(1 + \phi Q)}(0) \right]$$

$$= 0$$

Next, we show that the second term equals zero:

$$M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] \beta_j$$

$$= M \left[Q \sum_{q=1}^{Q} (I\{2 > q\}\beta_2 + \ldots + I\{J > q\}\beta_J) + Q(\beta_2 + \ldots + \beta_J) - Q \sum_{j=1}^{J} j\beta_j \right]$$

$$= M \left[Q \sum_{q=1}^{Q} (I\{2 > q\}\beta_2 + \ldots + I\{J > q\}\beta_J) + Q(\beta_2 + \ldots + \beta_J) - Q \sum_{j=1}^{J} j\beta_j \right]$$

$$= MQ[(1\beta_2 + 2\beta_3 + \ldots + Q\beta_J) + (\beta_2 + \ldots + \beta_J) - (2\beta_2 + \ldots + J\beta_J)]$$

$$= 0$$

Thus, the first two terms each equal zero, and so we can write:

$$E[\hat{\delta}] = M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[QI\{j > q\} - j + 1 + \frac{\phi Q(2q - Q - 1)}{2(1 + \phi Q)} \right] I\{j > q\}\delta_{j-q}$$

$$= M \sum_{q=1}^{Q} \sum_{j=1}^{J} \left[Q + 1 - \frac{\phi Q(Q+1)}{2(1 + \phi Q)} - j + \frac{\phi q Q}{1 + \phi Q} \right] I\{j > q\}\delta_{j-q}$$

$$= M \left[\frac{(2 + \phi Q)(Q+1)}{2(1 + \phi Q)} \sum_{q=1}^{Q} \sum_{j=1}^{J} I\{j > q\}\delta_{j-q} - \sum_{q=1}^{Q} \sum_{j=1}^{J} jI\{j > q\}\delta_{j-q} + \frac{\phi Q}{1 + \phi Q} \sum_{q=1}^{Q} \sum_{j=1}^{J} qI\{j > q\} \right]$$

$$\} \delta_{j-q} \Bigg]$$

Next, we simplify the three sums in the equation above as follows:

$$\sum_{q=1}^{Q} \sum_{j=1}^{J} I\{j > q\}\delta_{j-q} = \sum_{j=1}^{Q} (Q - j + 1)\delta_j$$

$$\sum_{q=1}^{Q} \sum_{j=1}^{J} jI\{j > q\}\delta_{j-q} = \sum_{j=1}^{Q} \left[\frac{(Q+1)(Q+2) - j(j+1)}{2} \right] \delta_j$$

$$\sum_{q=1}^{Q} \sum_{j=1}^{J} qI\{j > q\}\delta_{j-q} = \sum_{j=1}^{Q} \frac{(Q - j + 1)(Q - j + 2)}{2} \delta_j$$

Plugging these identities into the equation above, we obtain:

$$E[\hat{\delta}] = M \left[\frac{(2 + \phi Q)(Q+1)}{2(1 + \phi Q)} \sum_{j=1}^{Q} (Q - j + 1)\delta_j - \sum_{j=1}^{Q} \left[\frac{(Q+1)(Q+2) - j(j+1)}{2} \right] \delta_j \right.$$

$$+ \frac{\phi Q}{1 + \phi Q} \sum_{j=1}^{Q} \frac{(Q - j + 1)(Q - j + 2)}{2} \delta_j \Bigg]$$

$$= \frac{M}{2(1 + \phi Q)} \sum_{j=1}^{Q} [(j - Q - 1)(j + 2\phi Qj - Q(1 + \phi + \phi Q))]\delta_j$$

$$= \sum_{j=1}^{Q} \frac{6(j - Q - 1)((1 + 2\phi Q)j - (1 + \phi + \phi Q)Q)}{Q(Q+1)(\phi Q^2 + 2Q - \phi Q - 2)} \delta_j$$

This completes the proof.

E TABULAR RESULTS

Below, we have included the results from Figures 3 and 4 in tabular form, as well as Monte Carlo errors corresponding to Figure 3.

TABLE E1

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,J-1)}$) and LTE (Ψ_{J-1}) using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC). See Figure 3 for results in graphical form.

		(a) Instantaneous		(b) Lagged		(c) Curved		(d) Partially convex	
		TATE	LTE	TATE	LTE	TATE	LTE	TATE	LTE
Bias (%)	IT	−0.8	−0.8	−112.1	−108.1	−36.2	−46.2	−99.0	−99.3
	ETI	0.1	0.1	0.0	−0.1	−0.5	−1.7	1.2	1.4
	NCS-4	−4.2	−9.7	−11.1	−12.5	−0.7	−1.5	−3.4	−8.9
	MEC	9.1	18.4	21.0	21.3	−6.6	3.3	3.2	−3.1
Coverage (%)	IT	94.7	94.7	0.0	0.0	27.3	2.3	0.0	0.0
	ETI	94.1	95.2	94.9	94.6	95.0	94.8	95.7	94.6
	NCS-4	94.5	93.8	93.6	92.9	94.1	93.8	94.6	93.6
	MEC	92.6	93.5	91.1	93.5	93.3	95.3	93.9	91.4
MSE	IT	0.3	0.3	14.3	29.6	2.7	5.7	10.7	25.0
	ETI	0.8	3.1	0.8	3.3	0.8	3.1	0.8	3.2
	NCS-4	0.8	3.3	0.9	3.3	0.8	3.2	0.8	3.1
	MEC	0.7	1.6	0.8	1.8	0.7	1.1	0.7	1.7

TABLE E2

Simulation results: average pointwise mean squared error for the estimation of the entire effect curve using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC). See Figure 4 for results in graphical form.

	(a) Instantaneous	(b) Lagged	(c) Curved	(d) Partially convex
IT	0.003	0.199	0.035	0.143
ETI	0.013	0.014	0.013	0.014
NCS-4	0.014	0.017	0.013	0.013
MEC	0.008	0.010	0.009	0.013

TABLE E3

Monte Carlo standard errors corresponding to estimates in Table E1

		(a) Instantaneous		(b) Lagged		(c) Curved		(d) Partially convex	
		TATE	LTE	TATE	LTE	TATE	LTE	TATE	LTE
Bias (%)	IT	0.37	0.37	0.59	0.39	0.45	0.38	0.58	0.38
	ETI	0.56	1.12	0.85	1.14	0.67	1.10	0.86	1.13
	NCS-4	0.56	1.10	0.84	1.08	0.69	1.13	0.85	1.08
	MEC	0.43	0.53	0.52	0.52	0.59	0.64	0.79	0.82
Coverage (%)	IT	0.71	0.71	0.00	0.00	1.41	0.47	0.00	0.00
	ETI	0.75	0.68	0.70	0.71	0.69	0.70	0.64	0.71
	NCS-4	0.72	0.76	0.77	0.81	0.75	0.76	0.71	0.77
	MEC	0.83	0.78	0.90	0.78	0.79	0.67	0.76	0.89
MSE	IT	0.0001	0.0001	0.0004	0.0015	0.0006	0.0009	0.0012	0.0019
	ETI	0.0004	0.0014	0.0004	0.0015	0.0004	0.0013	0.0004	0.0014
	NCS-4	0.0004	0.0014	0.0004	0.0015	0.0003	0.0005	0.0004	0.0014
	MEC	0.0003	0.0007	0.0003	0.0007	0.0003	0.0005	0.0003	0.0008

Abbreviations:

0

CRT   cluster randomized trial

References

1. Hughes JP, Granston TS, Heagerty PJ. Current issues in the design and analysis of stepped wedge trials. Contemporary clinical trials 2015; 45: 55–60. [PubMed: 26247569]

2. Mdege ND, Man MS, Taylor CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. Journal of clinical epidemiology 2011; 64(9): 936–948. [PubMed: 21411284]

3. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Use of the stepped wedge design cannot be recommended: a critical appraisal and comparison with the classic cluster randomized controlled trial design. Journal of clinical epidemiology 2012; 65(12): 1249–1252. [PubMed: 22964070]

4. Mdege ND, Man MS, Taylor CA, Torgerson DJ. There are some circumstances where the stepped-wedge cluster randomized trial is preferable to the alternative: no randomized trial at all. Response to the commentary by Kotz and colleagues. Journal of clinical epidemiology 2012; 65(12): 1253. [PubMed: 22968178]

5. Kotz D, Spigt M, Arts IC, Crutzen R, Viechtbauer W. Researchers should convince policy makers to perform a classic cluster randomized controlled trial instead of a stepped wedge design when an intervention is rolled out. Journal of clinical epidemiology 2012; 65(12): 1255. [PubMed: 22964069]

6. Hemming K. Stepped wedge cluster randomized trials are efficient and provide a method of evaluation without which some interventions would not be evaluated. Journal of clinical epidemiology 2013; 66(9): 1058. [PubMed: 23849737]

7. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. Statistics in medicine 2015; 34(2): 181–196. [PubMed: 25346484]

8. Kotz D. The stepped wedge design does not inherently have more power than a cluster randomized controlled trial. Journal of clinical epidemiology 2013; 66(9): 1059. [PubMed: 23953488]

9. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC medical research methodology 2006; 6(1): 54. [PubMed: 17092344]

10. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemporary clinical trials 2007; 28(2): 182–191. [PubMed: 16829207]

11. Hopper A, Jamison M, Lewis W. Learning curves in surgical practice. Postgraduate medical journal 2007; 83(986): 777–779. [PubMed: 18057179]

12. Nickless A, Voysey M, Geddes J, Yu LM, Fanshawe TR. Mixed effects approach to the analysis of the stepped wedge cluster randomised trial—Investigating the confounding effect of time through simulation. PloS one 2018; 13(12): e0208876. [PubMed: 30543671]

13. Haugen AS, Søfteland E, Almeland SK, et al. Effect of the World Health Organization checklist on patient outcomes: a stepped wedge cluster randomized controlled trial. Annals of surgery 2015; 261(5): 821–828. [PubMed: 24824415]

14. Menon P, Nguyen PH, Saha KK, et al. Impacts on breastfeeding practices of at-scale strategies that combine intensive interpersonal counseling, mass media, and community mobilization: results of cluster-randomized program evaluations in Bangladesh and Viet Nam. PLoS medicine 2016; 13(10): e1002159. [PubMed: 27780198]

15. White EE, Downey J, Sathananthan V, et al. A community health worker intervention to increase childhood disease treatment coverage in rural Liberia: a controlled before-and-after evaluation. American journal of public health 2018; 108(9): 1252–1259. [PubMed: 30024811]

16. Granston T. Addressing lagged effects and interval censoring in the stepped wedge design of cluster randomized clinical trials. PhD thesis. 2014.

17. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials 2017; 18(1): 101. [PubMed: 28259174]

18. Matthews J, Forbes AB. Stepped wedge designs: insights from a design of experiments perspective. Statistics in Medicine 2017; 36(24): 3772–3790. [PubMed: 28786236]

19. Kasza J, Hemming K, Hooper R, Matthews J, Forbes A. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. Statistical methods in medical research 2019; 28(3): 703–716. [PubMed: 29027505]

20. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. Statistical Methods in Medical Research 2020: 0962280220932962.

21. Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. Statistics in medicine 2016; 35(13): 2149–2166. [PubMed: 26748662]

22. Hooper R, Teerenstra S, Hoop dE, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. Statistics in medicine 2016; 35(26): 4718–4728. [PubMed: 27350420]

23. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. Statistics in medicine 2019; 38(11): 1918–1934. [PubMed: 30663132]

24. Friedman J, Hastie T, Tibshirani R, others . The elements of statistical learning. 1. Springer series in statistics New York . 2001.

25. Shively TS, Sager TW, Walker SG. A Bayesian approach to non-parametric monotone function estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2009; 71(1): 159–175.

26. Kenny A, Wolock C. SimEngine: an R package for structuring statistical simulations. 2021.

27. Haines TP, Bowles KA, Mitchell D, et al. Impact of disinvestment from weekend allied health services across acute medical and surgical wards: 2 stepped-wedge cluster randomised controlled trials. PLoS medicine 2017; 14(10): e1002412. [PubMed: 29088237]

28. Golden MR, Kerani RP, Stenger M, et al. Uptake and population-level impact of expedited partner therapy (EPT) on Chlamydia trachomatis and Neisseria gonorrhoeae: the Washington State community-level randomized trial of EPT. PLoS Med 2015; 12(1): e1001777. [PubMed: 25590331]

29. De Chaisemartin C, D'Haultfœuille X. Double fixed effects estimators with heterogeneous treatment effects. tech. rep., Working paper; 2016.

30. Callaway B, Sant'Anna PH. Difference-in-differences with multiple time periods. Journal of Econometrics 2020.

31. Kasza J, Forbes AB. Information content of cluster–period cells in stepped wedge trials. Biometrics 2019; 75(1): 144–152. [PubMed: 30051909]

32. Kasza J, Taljaard M, Forbes AB. Information content of stepped-wedge designs when treatment effect heterogeneity and/or implementation periods are present. Statistics in Medicine 2019; 38(23): 4686–4701. [PubMed: 31321806]

33. Samuel-Cahn E Combining unbiased estimators. The American Statistician 1994; 48(1): 34–36.

34. Scott JM, deCamp A, Juraska M, Fay MP, Gilbert PB. Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. Statistical methods in medical research 2017; 26(2): 583–597. [PubMed: 25267551]

35. Kennedy-Shaffer L, De Gruttola V, Lipsitch M. Novel methods for the analysis of stepped wedge cluster randomized trials. Statistics in Medicine 2020; 39(7): 815–844. [PubMed: 31876979]

36. Thompson J, Davey C, Fielding K, Hargreaves J, Hayes R. Robust analysis of stepped wedge trials using cluster-level summaries within periods. Statistics in medicine 2018; 37(16): 2487–2500. [PubMed: 29635789]

37. Hughes JP, Heagerty PJ, Xia F, Ren Y. Robust inference for the stepped wedge design. Biometrics 2020; 76(1): 119–130. [PubMed: 31237680]

38. Zhou X, Liao X, Spiegelman D. "Cross-sectional" stepped wedge designs always reduce the required sample size when there is no time effect. Journal of clinical epidemiology 2017; 83: 108. [PubMed: 28093263]

39. Inc. WR. Mathematica, Version 12.3.1. Champaign, IL, 2021.

40. Voldal EC, Xia F, Kenny A, Heagerty PJ, Hughes JP. Random effect misspecification in stepped wedge designs. Clinical Trials: 17407745221084702.

41. Bowden R, Forbes AB, Kasza J. Inference for the treatment effect in longitudinal cluster randomized trials when treatment effect heterogeneity is ignored. Statistical Methods in Medical Research 2021; 30(11): 2503–2525. [PubMed: 34569853]
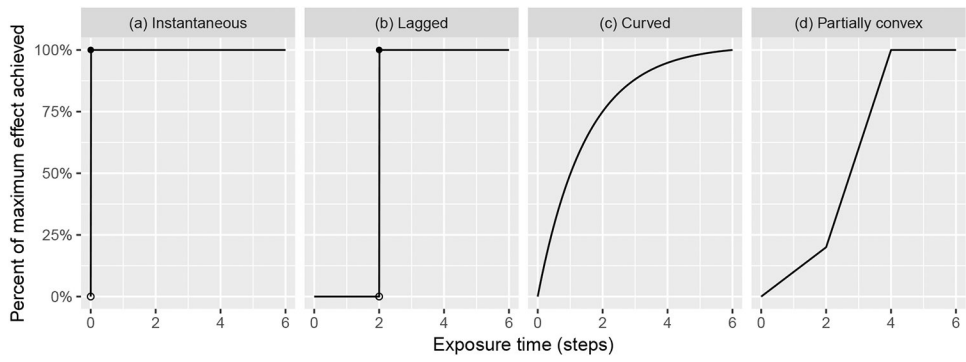
FIGURE 1.

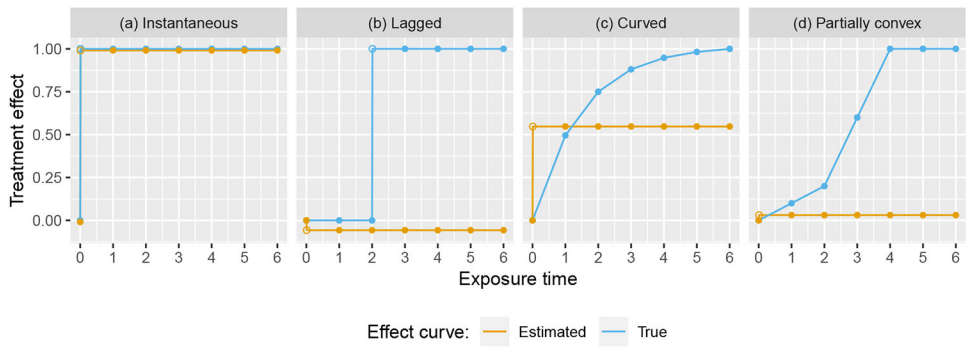Several possible effect curves: treatment effect as a function of exposure time

FIGURE 2.

Four possible true effect curves plotted against the expected effect curves estimated from an IT model, for a design with $Q = 6$ sequences and $\phi = 0.5$
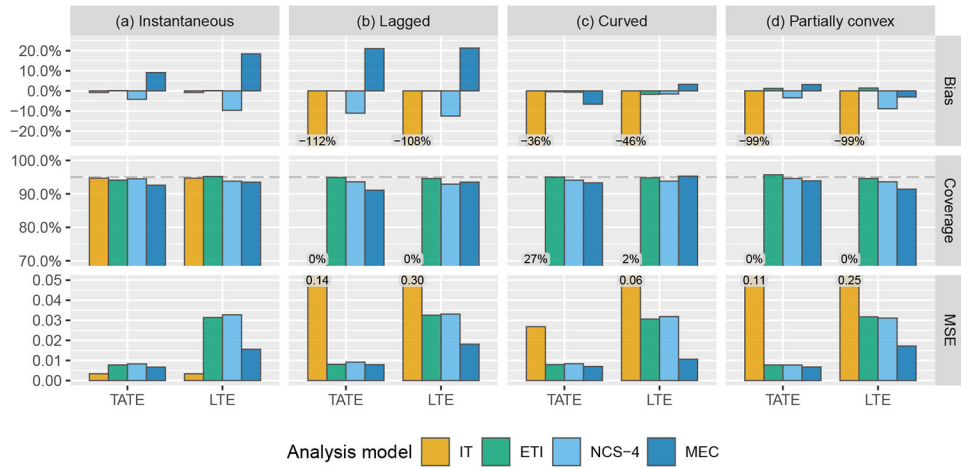
FIGURE 3.

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,J-1]}$) and LTE (Ψ_{J-1}) using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC). Numeric values displayed over graph bars represent the height of the bars that are cut off because of the scale of the axes; see Table E1 for results in tabular form.
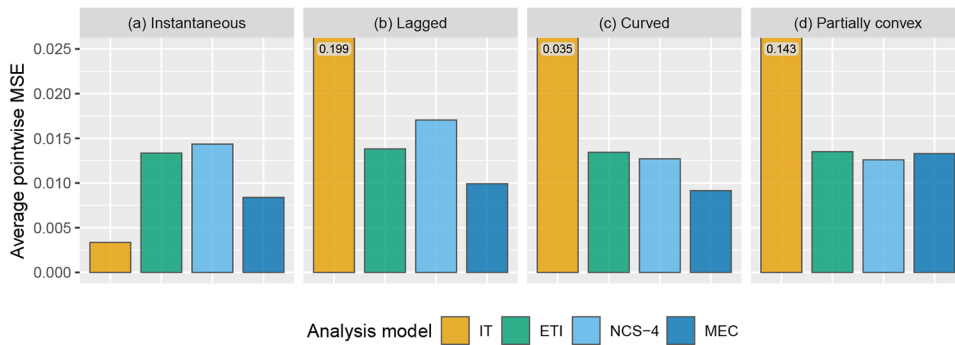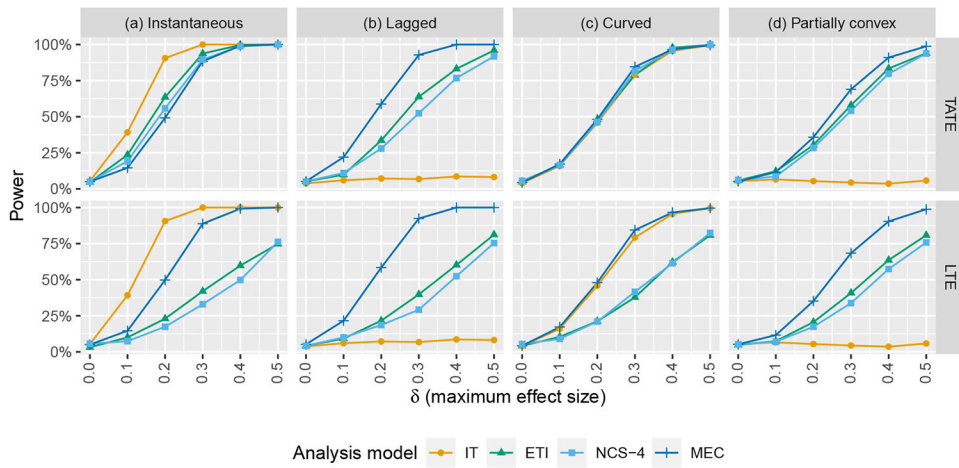
FIGURE 4.

Simulation results: average pointwise mean squared error for the estimation of the entire effect curve using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC). Numeric values displayed over graph bars represent the height of the bars that are cut off because of the scale of the axes; see Table E2 for results in tabular form.

FIGURE 5.

Simulation results: power of Wald-type hypothesis tests for testing null hypotheses related to the TATE ($\Psi_{(0,J-1]} = 0$) and the LTE ($\Psi_{J-1} = 0$) using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC)
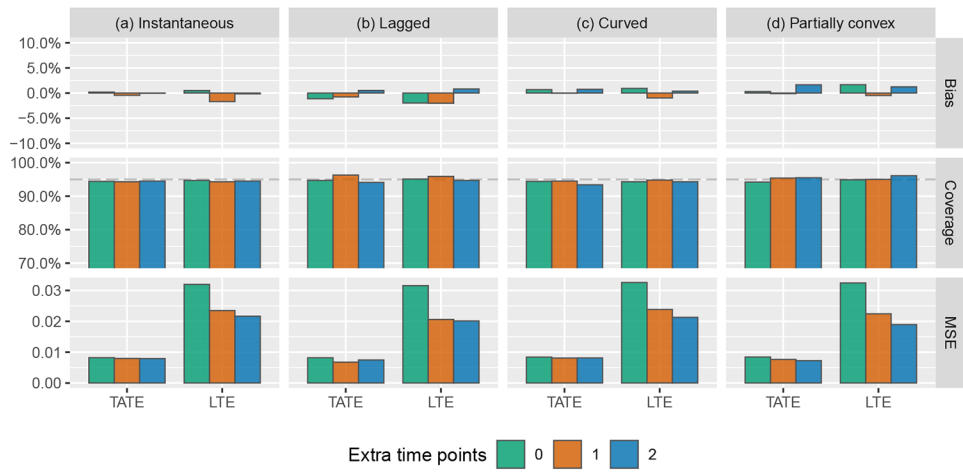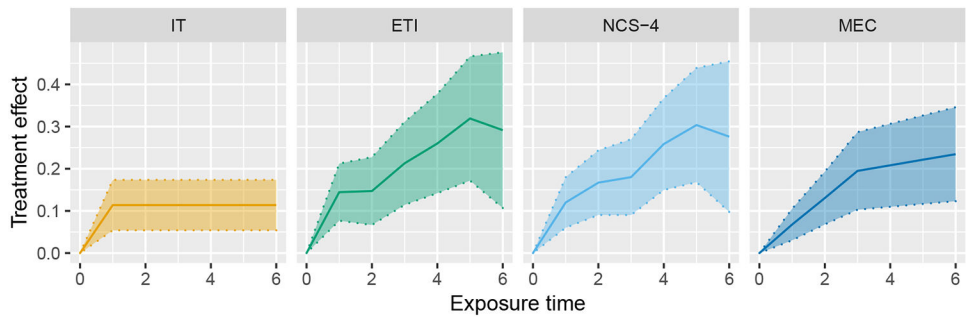
FIGURE 6.

Simulation results: bias, coverage, and mean squared error (MSE) for the estimation of the TATE ($\Psi_{(0,j-1]}$) and LTE (Ψ_{j-1}) using the exposure time indicator (ETI) model with 0, 1, or 2 extra time points added to the end of the study
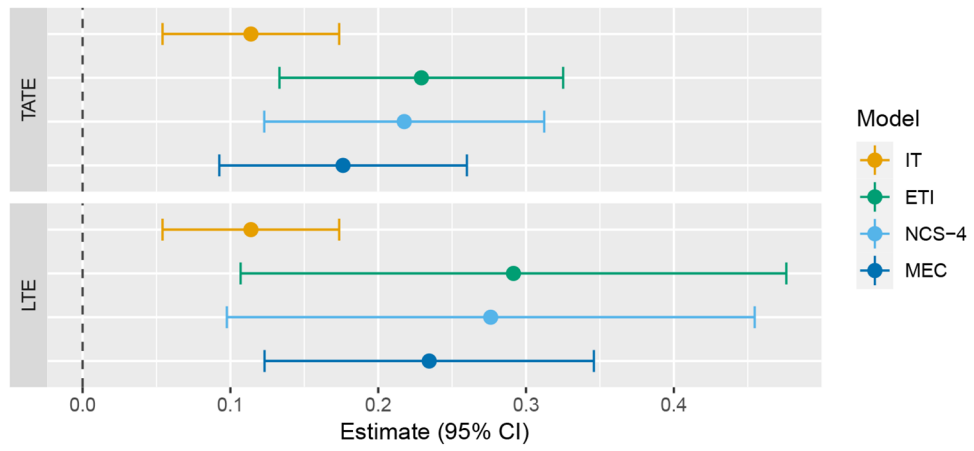
FIGURE 7.

Forest plot of TATE and LTE estimates from the Australia disinvestment trial using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC)

FIGURE 8.
Estimation of the effect curve from the Australia disinvestment trial using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC)
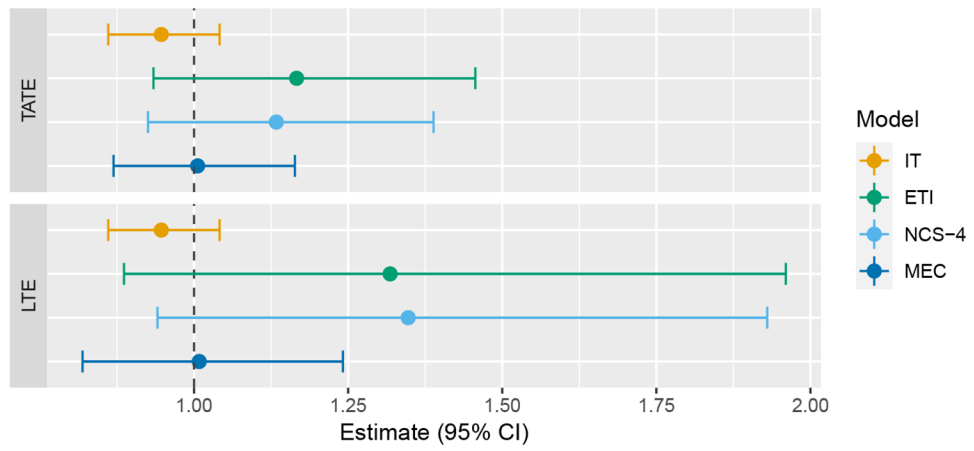
FIGURE 9.
Forest plot of TATE and LTE estimates from the WA State EPT Trial (odds ratios) using
the following four models: immediate treatment effect (IT), exposure time indicator (ETI),
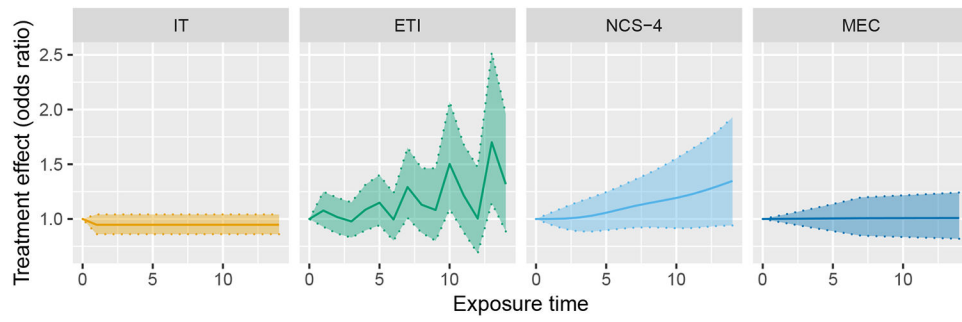natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC)

FIGURE 10.
Estimation of the effect curve from the WA State EPT Trial using the following four models: immediate treatment effect (IT), exposure time indicator (ETI), natural cubic spline with 4 degrees of freedom (NCS-4), monotone effect curve (MEC)