Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 51-66

Advances in Stepped Wedge Design: A Comprehensive Review

Soumadeb Pain and Satya Prakash Singh

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, 208016, India

Received: 12 June 2025; Revised: 23 June 2025; Accepted: 26 June 2025

Abstract

Stepped wedge designs (SWDs) are increasingly gaining popularity in cluster randomized trials. This review provides a comprehensive overview of stepped wedge cluster randomised trials (SW–CRTs), beginning with their historical development and rationale in the Introduction. We classify and compare different types of stepped wedge designs, highlighting their relative advantages and practical considerations. We then examine the primary statistical models used for analysis, key approaches to sample size determination, and the impact of various intracluster and temporal correlation structures on trial inference. Special attention is given to trials with unequal cluster sizes, addressing design adaptations and efficiency implications. We review recent advances in Bayesian optimal design strategies for SW–CRTs and extend the discussion to include adaptations for non-normal outcome data. As an alternative framework, we explore the staircase design, comparing its logistical and analytical features with those of traditional stepped wedge trials.

Key words: Staircase design; Stepped wedge trials; Cluster randomized trials; Optimal design; Bayesian design.

1. Introduction

Cluster Randomized Trials (CRTs) are pivotal in evaluating interventions where groups, rather than individuals, are randomized. Statistical methods for CRTs have been the focus of extensive research over the past several decades and are well-documented in various methodological reviews [Donner and Klar (2000), Turner *et al.* (2017a), and Turner *et al.* (2017b)]. Among CRT designs, the stepped wedge cluster randomized trial (SW-CRT) has garnered increasing attention, alongside traditional parallel and crossover CRTs [Brown and Lilford (2006), Mdege *et al.* (2011), Hemming *et al.* (2015)]. While parallel designs randomize clusters to fixed intervention or control arms and crossover designs alternate clusters between arms over time, the SW-CRT employs a unidirectional roll-out of all the clusters from control to intervention in sequential "steps". A SW–CRT comparing control and intervention conditions is illustrated in Table 1. The order in which the different individuals

Corresponding Author: Satya Prakash Singh
Email: singhsp@iitk.ac.in

or clusters receive the intervention is random. The study continues until all the clusters are assigned to the intervention and outcome data will be collected from each cluster. The SW design offers a versatile framework for CRTs, particularly in public health and service delivery. SW–CRTs offer several distinctive advantages including:

- **Evaluation During Rollout:** SW–CRTs are particularly useful for assessing the community level effectiveness of an intervention while it is being gradually rolled out across clusters.

- **Acceptability (Social, Political, Ethical):** Since all clusters eventually receive the intervention, this design is often more acceptable to stakeholders, especially in contexts where withholding a potentially beneficial intervention may be controversial.

- **Logistical and Financial Feasibility:** In many cases, it is not practical due to resource, personnel, or policy constraints to implement the intervention across all units simultaneously. SW–CRTs accommodate such staged implementation.

- **Statistical Efficiency:** Because each cluster serves as its own control at different time points, this design can increase statistical efficiency and may require fewer clusters compared to parallel-arm trials.

While its ethical and practical benefits are significant, researchers must address analytical complexities and potential biases. Hemming and Taljaard (2020) discussed several key factors that should be considered when implementing a SW-CRT.

The Gambia Hepatitis Intervention Study (Hall *et al.* (1987)) is the first ever reported stepped wedge trial which is also the longest running. The study was set up in 1986 to investigate whether vaccination against hepatitis B in infancy could reduce the risk of liver cancer over the next 30 to 40 years of life. The usefulness of SW-CRTs was recognized only later, but they are now highly regarded and widely used in medical research, as demonstrated by a recent review by Varghese *et al.* (2025), which examines studies published in high-impact journals. Some previous reviews on SW–CRTs include Brown and Lilford (2006) and Beard *et al.* (2015).

This paper focuses on reviewing recent statistical developments related to the design of SW-CRT trials. In doing so, we also introduce several variants of SW-CRTs that are widely used in current practice. The organization of this article is as follows: Section 2 reviews types of stepped wedge designs; Section 3 covers models, sample size determination and correlation structures; Section 4 examines unequal cluster sizes; Section 5 presents Bayesian optimal SW designs; Section 6 addresses non-normal outcome data; Section 7 introduces the staircase design as an alternative to SWD; and Section 8 provides discussion and conclusions.

2. Types of stepped wedge designs

Following (Copas *et al.*, 2015), SW–CRTs can be broadly classified in three categories: (i) Cohort SW–CRTs; (ii) Cross–sectional; and (iii) Continuous recruitment SW–CRTs.

2.1. Cohort SW–CRTs

When researchers track a group of individuals over time and assess their health outcomes at regular intervals, this is referred to as a cohort study design. A cohort design further classified in two categories: (a) Closed cohort SW–CRTs and (b) Open cohort SW–CRTs. In a stepped wedge trial, where the same clusters are revisited at different time points, it is common to measure outcomes for some or all of the same individuals on multiple occasions some before and some after the intervention is introduced. When the measurements are taken from same individuals at each time point in a cluster, the design is known as a closed cohort stepped wedge trial (Li *et al.* (2018a), Li (2020), Gasparini *et al.* (2025)). While this is a reasonably appropriate design, it may not always be realistic. In practice, it is more likely that participants may enter or leave the study over time. So, when some individuals are the same and others differ across measurement periods within a cluster, the design is referred to as an open cohort stepped wedge trial (Copas *et al.* (2015), Kasza *et al.* (2020)). Here, the term "open cohort" reflects the natural flow of participants in and out of the study population.

2.2. Cross–sectional SW–CRT

Sometimes clusters that are very large or densely populated, where researchers do not try to measure outcomes for everyone, but instead select a small, random sample at each visit. In that case, the chance of observing the same person twice is minimal. At this point, no individual is followed longitudinally across multiple time points (steps). Instead, each sample taken from a cluster represents a snapshot or cross-section of the population at that specific time. This type of design is known as a repeated cross-section stepped wedge trial (Hussey and Hughes (2007), Martin *et al.* (2019), Thompson *et al.* (2017)). It is relatively logistically simpler than cohort SW, since we do not need to track the same individuals over months, or in other words, we do not have to deal with individual's auto correlation.

2.3. Continuous recruitment SW–CRTs

In aforementioned SW–CRTs, we have described scenarios where extending the duration of a cluster randomized trial involves returning to the same clusters multiple times to collect outcome data. But what if, instead, participants enter and exit the trial in a continuous flow like an ongoing stream of eligible individuals? In this case, extending the trial simply means recruiting over a longer period of time, allowing more people from that continuous stream to be included. This type of design is known as a continuous recruitment stepped wedge trial (Hooper and Copas (2019), Hooper *et al.* (2020)). An example of a this type of stepped wedge trial is the Gambia Hepatitis Intervention Study discussed in the Introduction section. In this study, new eligible participants (newborn infants) arrived at a fairly steady rate, as is natural. In any continuous one year period of recruitment the researchers expected to recruit around 30,000 children into the study, and by scheduling the trial over a total of four years they hoped to see 120,000.

Table 1: A T–period stepped wedge design with $T-1$ sequences, comparing the control (in grey) and treatment (in blue).

Sequence	Time					
	0	**1**	**2**	**3**	**4**	**5**
1						
2						
3						
4						
5						

Cluster unexposed to intervention Cluster exposed to intervention

3. Models, sample size determination and correlation structures in SW–CRTs

This section reviews key developments in the statistical methodology of SW-CRTs. We begin by exploring various modeling structures, followed by a review of sample size determination methods and the different forms of correlation, a key aspect of SW–CRTs.

3.1. Models

Broadly, SW–CRTs are analyzed using conditional (cluster/subject/time specific) and marginal (population average) models. Conditional models, commonly implemented via linear mixed effects models (LMMs) or generalized linear mixed-effects models (GLMMs), account for clustering through random effects and estimate intervention effects conditional on these latent cluster/subject/time–level factors. Conditional models are utilized by Hussey and Hughes (2007), Hughes *et al.* (2015), Hooper *et al.* (2016), Girling and Hemming (2016), Kasza *et al.* (2019), Kasza and Forbes (2019), and Hemming *et al.* (2018).

Marginal models, typically fitted via generalized estimating equations (Liang and Zeger (1986)), directly target the population average treatment effect, offering robust inference even under correlation structure misspecification. Within each framework, a variety of correlation structures such as exchangeable, nested exchangeable, and exponential decay have been proposed to capture within and between period intraclass correlations (ICCs). Marginal models are used by Hussey and Hughes (2007), Li *et al.* (2018b), Ford and Westgate (2020), Li (2020), Thompson *et al.* (2021), and Li *et al.* (2022).

The choice between conditional and marginal approaches affects both interpretation and efficiency. There has existed controversy about the use of marginal and conditional models. Lee and Nelder (2004) discussed the advantages of conditional models over marginal models and regarded the conditional model as fundamental, from which marginal predictions can be made. Various models employed in SW design are thoroughly discussed in the review paper by Li and Wang (2022).

3.2. Sample size determination

3.2.1. Foundational work

A critical aspect of SWDs is sample size calculation, which has seen significant methodological advancements. The seminal paper by Hussey and Hughes (2007) introduced analytical formulas for power and sample size calculations in SWDs. The model proposed by Hussey and Hughes (2007), though foundational for sample size estimation in SWDs, has notable limitations. First, it assumes a cross-sectional design with no repeated measurements on individuals, rendering it unsuitable for cohort-based studies where individual correlation must be accounted for (Hooper *et al.*, 2016). Second, it presumes constant cluster sizes and a simplistic intracluster correlation (ICC) structure, ignoring variability in cluster sizes (Matthews, 2020) or more complex correlation patterns (*e.g.*, decaying correlations over time). Third, the model assumes fixed time effects and a constant intervention effect, failing to accommodate time-varying treatment effects or interactions between time and intervention exposure (Kenny *et al.*, 2022). Finally, it is restricted to continuous outcomes and does not generalize readily to binary, count, or survival data without modification (Zhou *et al.*, 2020).

3.2.2. Sample size calculations based on design effects

The design effect quantifies the increase in variance of an estimator due to deviations from a simple random sampling design. In cluster-based studies, it accounts for correlations within clusters, which reduces the effective sample size. The standard approach to calculating sample size in parallel group CRTs begins with estimating the required sample size under individual randomization, denoted as N_u. This unadjusted sample size is then scaled by the design effect $[1+(n-1)\rho]$ to account for clustering, where n is the number of individuals per cluster and ρ is the intracluster correlation coefficient (Donner and Klar, 2000). To adopt a similar framework, Woertman *et al.* (2013) derived the following design effect for SWDs:

$$DE_{sw} = \frac{1+\rho(ktn+bn-1)}{1+\rho\left(\frac{1}{2}ktn+bn-1\right)}\frac{3(1-\rho)}{2t\left(k-\frac{1}{k}\right)}.$$

Here, k represents the number of steps, b is the number of baseline measurements, and t is the number of measurements after each step. Thus, each cluster is measured $(b+kt)$ times. This design effect appropriately adjusts for both clustering and the stepped wedge structure and the required sample size for a stepped wedge trial is $N_{sw} = N_u DE_{sw}$.

The design effect DE_{sw} is influenced by three key parameters: the number of post–step measurements t, the number of baseline measurements b, and the number of steps k. Increasing any of these reduces the design effect and, consequently, the required sample size. In contrast, increasing the cluster size n slightly increases the design effect. Additionally, DE_{sw} depends on the intracluster correlation coefficient (ICC), ρ, which reflects variability between clusters. While ρ is contextd ependent and not under direct control, it should be estimated using prior studies, pilot data, or domain knowledge. As ρ increases, the design effect initially rises and then begins to decline. Woertman *et al.* (2013) have shown that increasing the number of steps improves efficiency in terms of sample size and also the gain is substantially larger when increasing from 2 to 3 steps than from 6 to 12 steps.

3.2.3. Simulation–based sample size calculations

Analytical sample size formulas, while computationally efficient, are often constrained by simplifying assumptions that limit their applicability in real world SW–CRTs. For example, Hussey and Hughes (2007), Woertman *et al.* (2013) *etc* assumed balanced design and intervention effect is modeled as constant across clusters. Also analytical formulas work well for continuous outcomes but struggle with binary or count outcomes Xia *et al.* (2021) or when repeated measures are taken on the same individuals over time, due to the additional level of correlation implied in this case. Simulation–based sample size calculation has emerged as a robust and flexible approach for designing SW-CRT, particularly when analytical formulas are insufficient due to the complexity of the design or outcome types. Simulation-based sample size calculation typically follows these steps:

- Define the Data-Generating Model: Specify fixed effects (*e.g.*, intervention effect, time trends) and random effects (*e.g.*, cluster-level or time-level variability).
- Simulate Datasets: Generate repeated datasets under the assumed model, incorporating design parameters (for example number of clusters, steps, and observations per cluster).
- Analyze Simulated Data: Apply the planned statistical method (*e.g.*, mixed-effects regression) to each dataset, estimate the intervention effect and its standard error and p-value and then record the proportion of simulations where the intervention effect is statistically significant (empirical power).
- Iterate Until Target Power is Achieved: Adjust parameters (*e.g.*, cluster size or number of steps) and repeat simulations until the desired power is reached

But at the same time there are challenges in simulation-based sample size calculation. Simulations require significant computational resources, especially for large trials or complex models. Analyzing thousands of datasets with complex models can be slow. Parallel computing is often essential and also complex models may fail to converge in some simulations. Further, results depend on accurate pre–specification of nuisance parameters (*e.g.*, ICC), which may be uncertain in practice.

3.3. Correlation structures

Stepped wedge designs inherently involve longitudinal and clustered data, leading to multiple correlation structures that complicate statistical analysis. These structures arise from repeated measurements within clusters over time, participant-level dependencies in cohort designs, and temporal trends.

3.3.1. Correlation parameters in SW–CRTs

Hemming *et al.* (2015) incorporated both within-period and between-period ICCs in their sample size calculation for cross-sectional designs. Hooper *et al.* (2016) and Li *et al.* (2018b) extended this by considering a three correlation structure that also accounts for within individual repeated measurements in closed cohort designs.

- Within period intracluster correlation (wp–ICC): Measures similarity of outcomes within the same cluster and time period.

- Between period intracluster correlation (bp–ICC): Captures correlation between outcomes from the same cluster across different periods.

- Individual level autocorrelation: Relevant in closed cohort designs where the same participants are measured repeatedly.

For example, in cross sectional designs (new participants each period), wp–ICC and bp–ICC dominate, while closed cohort designs require an additional parameter for individual autocorrelation. Ignoring these distinctions can lead to biased variance estimates and underpowered studies (Girling and Hemming, 2016).

Table 2: Different types of correlation structures in SW design ($0 < r < 1$ is any constant value)

(a) Constant ICC over time: within period ICC = between period ICC
Used in Hussey and Hughes (2007) − no decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	ρ	ρ	ρ	ρ
Period 2		ρ	ρ	ρ	ρ
Period 3			ρ	ρ	ρ
Period 4				ρ	ρ
Period 5					ρ

(b) Fixed between period ICC and within period ICC > between period ICC
Used in Hooper *et al.* (2016)− no decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	$r\rho$	$r\rho$	$r\rho$	$r\rho$
Period 2		ρ	$r\rho$	$r\rho$	$r\rho$
Period 3			ρ	$r\rho$	$r\rho$
Period 4				ρ	$r\rho$
Period 5					ρ

(c) Between ICCs decay exponentially and within period ICC > between period ICC
Used in Kasza and Forbes (2019)− allows decay

	Period 1	Period 2	Period 3	Period 4	Period 5
Period 1	ρ	$r\rho$	$r^2\rho$	$r^3\rho$	$r^4\rho$
Period 2		ρ	$r\rho$	$r^2\rho$	$r^3\rho$
Period 3			ρ	$r\rho$	$r^2\rho$
Period 4				ρ	$r\rho$
Period 5					ρ

3.3.2. Modeling decaying correlation structures

In longitudinal studies, "decay in correlation over time" refers to the phenomenon where the correlation between two measurements decreases as the time interval between them increases. In other words, observations made closer together in time tend to be more similar (more correlated) than observations that are farther apart.

The assumption of constant ICC, as used by Hooper *et al.* (2016) and Li *et al.* (2018b), may not reflect real world data structures. Therefore, alternative design and analysis strategies that account for temporal correlation decay are essential in stepped wedge trials. In cross sectional designs, where different individuals are observed in each period, some studies (Hemming *et al.*, 2015) allowed between period ICC to differ from within period ICC but assumed constancy across time. Kasza *et al.* (2019), Kasza and Forbes (2019) introduced a nonuniform correlation model incorporating exponential decay, improving sample size estimation. Grantham *et al.* (2019) extended this to continuous time correlation decay in multiple periods CRTs with continuous recruitment. Ignoring correlation decay, as shown by Kasza *et al.* (2019), can lead to misestimate intervention effects and incorrect sample size calculations.

4. SW–CRTs with unequal cluster size

Methods for calculating power and sample size in SW-CRTs assuming equal cluster sizes have been extensively discussed in the literature see for example Hussey and Hughes (2007), Woertman *et al.* (2013), Baio *et al.* (2015), and Hemming and Taljaard (2016). In many studies, such as observational studies, unequal cluster sizes are a common occurrence. This presents significant challenges in the design and analysis of SW–CRTs. A comprehensive methodological review addressing unequal cluster sizes in cluster randomized trials, including SW-CRTs, is provided in Zhan *et al.* (2021b). The impact of cluster size imbalance on the power is discussed in Ouyang *et al.* (2020). Martin *et al.* (2019) examined how randomly allocating clusters of varying sizes to sequences impacts different aspects of the analysis. They investigated cluster–balanced stepped wedge designs (SWDs) with unequal cluster sizes and observed that, when the total number of individuals is fixed, such designs can be more efficient than those with equal cluster sizes. This finding contrasts with traditional cluster balanced designs, where equal sized clusters are typically considered optimal. Girling (2018) investigate the impact of unequal cluster size and found the expressions for the relative efficiency (RE) of the treatment effect estimate relative to that for the equal cluster design with the same total number of observations. Matthews (2020) proposed near optimal designs for unequal cluster size. Kristunas *et al.* (2017) proposed corrections to the design effect(DE) for SWD with unequal cluster sizes. Girling (2018) investigate the impact of unequal cluster size and found the expressions for the relative efficiency (RE) of the treatment effect estimate relative to that for the equal cluster design with the same total number of observations. Using simulations Martin *et al.* (2019) showed that the while the average power reduction in SW-CRTs is smaller than in parallel designs, the variance in power across allocations is higher, particularly with fewer clusters.

Typically, larger clusters are assigned to the extreme sequences. However, this pattern may not hold in closed-cohort stepped wedge designs (SWDs), where optimal allocation depends on various correlation parameters. In a working paper, we observed that an efficient

design tends to allocate an equal number of clusters to sequences i and $T-1-i$ for $i = 1, \ldots, T-1$. A similar symmetry is also observed in the allocation of total cluster size across these sequences. The determination of optimal cluster–to–sequence proportions in the context of unequal cluster sizes remains an area that requires further detailed investigation.

5. Bayesian optimal SWD

An optimal design is obtained by optimizing a specific criterion. For example, by minimizing the variance of the estimated treatment effect or by maximizing the study's power or precision. Lawrie *et al.* (2015) found out optimal allocation of clusters into sequences under the linear mixed effect model given by Hussey and Hughes (2007) by minimizing the variance of the treatment effect. They demonstrated that when cluster sizes are equal, the extreme sequences (first and last) receive the same level of allocation, while all intermediate sequences receive an equal but smaller allocation compared to the extremes. This work is then extended to closed cohort SW–CRT designs with repeated measures per subject by Li *et al.* (2018a). Thompson *et al.* (2017) examined the optimal structure of stepped wedge cluster randomized trial (SW–CRT) designs under the assumption of normally distributed data and equal allocation of clusters across sequences. In contrast, Zhan *et al.* (2018) explored optimal designs where some clusters may not be sampled during certain stages of the trial. Optimal design, thus obtained, is called locally optimized design as they are sensitive to the choice of different correlation parameters. More recently, to obtain a robust optimal design a Bayesian approach is adopted. Zhan *et al.* (2021a) demonstrated that incorporating prior information on time effects through a Bayesian approach can significantly reduce the required sample size. However, due to the risk of bias from mis-specified prior distributions, they do not recommend this as the default method for sample size calculation. Nevertheless, when it is difficult to recruit enough clusters or participants, using external information on time effects with a Bayesian approach can help assess if a smaller sample size would still be sufficient, making it easier to decide whether the trial can go ahead. Singh (2024) proposed a Bayesian optimal SWD by placing priors on the ICC and demonstrating robustness against ICC misspecification compared to locally optimal designs. Under a marginal (GEE) model with either exchangeable or exponential-decay working correlation, Etfer *et al.* (2024) developed a framework for finding Bayesian D-optimal SW designs for binary outcomes.

Bayesian designs for stepped wedge trials remain a significantly underdeveloped area of research with considerable potential. For instance, in closed–cohort studies, the presence of multiple correlation parameters introduces substantial uncertainty in the design process. This challenge can be effectively addressed by adopting a Bayesian framework. Moreover, in the case of non–normal responses where the optimal design criteria depend on unknown model parameters, a Bayesian approach can offer substantial advantages.

6. SW design for non–normal data

In recent years, a substantial body of work has extended the SW-CRT framework to accommodate non-normal outcomes, most notably binary and count data through a variety of methodological and practical innovations. Stepped-wedge trials with non-normal outcomes (counts or binary) extend the usual mixed-effects framework by replacing the linear mixed model with a generalized linear mixed model (GLMM) or generalized estimating equations (GEE). Broadly, for binary outcomes one uses logistic-link GLMMs or marginal GEE,

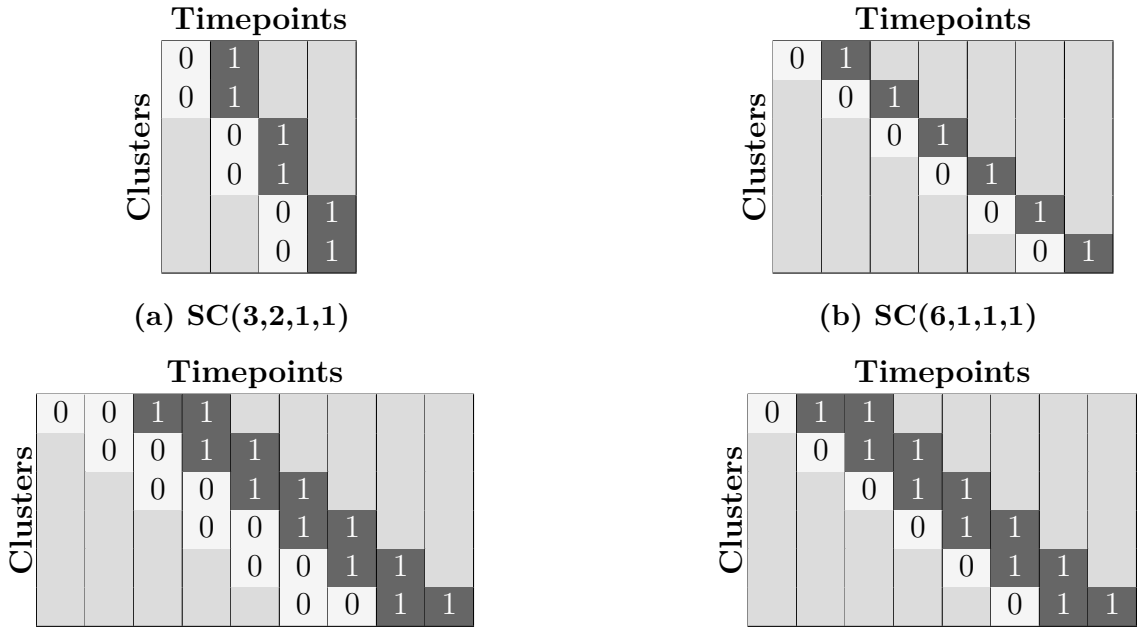
whereas for counts one adopts Poisson (or negative-binomial) GLMMs. Zhou *et al.* (2020) developed a numerical method for the power analysis for stepped-wedge cluster randomized trials (SW-CRTs) with binary outcomes, utilizing a maximum-likelihood estimation framework. Their approach allows researchers to assess the statistical power of complex SW-CRT designs without relying on simplified analytical approximations, making it particularly useful for settings with unequal cluster sizes or varying intraclass correlations.Wang *et al.* (2021) found out a sample size and power calculation method using GEE that can be broadly applied to both closed-cohort and cross-sectional SW-CRTs with binary outcomes. Also, they introduced a correction method to address the problem of underestimated variance in the GEE approach when the number of clusters is small in SW-CRTs. Building on the Laplace approximation of Breslow and Clayton (1993), Xia *et al.* (2021) have derived an analytical variance formula for the intervention effect estimator using GLMM, encompassing both normal (identity link) and non-normal (*e.g.*, logistic, Poisson) outcomes. Lastly as mentioned in the previous section, Etfer *et al.* (2024) develop a Bayesian D-optimal design framework for stepped-wedge cluster randomized trials with binary outcomes by combining generalized estimating equations and approximate design theory under both exchangeable and exponential-decay correlation structures.

7. Staircase design: An alternative to SWD

Stepped wedge designs require clusters to collect data across all trial periods, leading to high logistical and financial burdens. A staircase design is an "incomplete" variant of the stepped-wedge, in which each cluster contributes data only for a small number of periods immediately before and after its switch from control to intervention. The staircase design was first formalized by Grantham *et al.* (2024), who noted that the most informative observations in a stepped-wedge lie along its main diagonal (the "zigzag" of switches) and proposed focusing data collection there only. Like a stepped-wedge, all clusters eventually receive the intervention and the rollout is staggered; unlike a complete stepped-wedge, clusters do not collect data in every period, reducing burden and potentially attrition.

The general class of staircase designs is denoted by $SC(S, K, R_0, R_1)$, where S and R_0 denote the number of distinct treatment sequences and the number of clusters per sequence, R_0 is the number of control periods before the switch to intervention, and R_1 is the number of intervention periods after the switch. Different types of staircase designs can be achieved depending on the values of R_0 and R_1 (see Figure 1). In total, the design includes SK clusters, and the trial spans $S + R_0 + R_1 - 1$ periods. Clusters in sequence s are observed from period s through $s + R_0 + R_1 - 1$. A balanced staircase design has equal numbers of control and intervention periods in each sequence (*i.e.* $R_0 = R_1$). In contrast, an imbalanced staircase design allows for different numbers of pre and post switch periods ($R_0 \neq R_1$).

Grantham *et al.* (2024) have derived explicit expressions for the variance of the generalized least squares estimator of treatment effect for the basic staircase design under the assumption that the observed periods in each sequence follow the same schedule of control and intervention periods. This expression can be used to calculate sample size and power for staircase designs. Grantham *et al.* (2025) examined the relative efficiency of the stepped wedge design compared to various forms of the basic staircase design, where each sequence consists of one control period followed by one intervention period. Their analysis began with a basic staircase design embedded within a stepped wedge framework, and extends to



(a) SC(3,2,1,1) **(b) SC(6,1,1,1)**

(c) SC(6,1,2,2) **(d) SC(6,1,1,2)**

Figure 1: Design schematics for several staircase designs with 6 clusters: a basic staircase with two clusters assigned to each of three unique sequences (top left), a basic staircase with one cluster assigned to each of six unique sequences (top right), a balanced staircase with two control periods followed by two intervention periods in each sequence and one cluster assigned to each of six unique sequences (bottom left), and an imbalanced staircase with one control period followed by two intervention periods in each sequence and one cluster assigned to each of six unique sequences (bottom right).

versions with either more clusters or larger cluster-period sizes–some maintaining the same total number of participants as the stepped wedge design, and others using fewer participants overall. The relative efficiency of these designs is influenced by the intracluster correlation structure, correlation parameters, and trial configuration, including the number of sequences and the size of each cluster–period. They concluded that basic staircase design is a particularly lean and potentially powerful alternative to the stepped wedge design as across a broad range of realistic trial scenarios, the basic staircase design often provides greater statistical power than the stepped wedge design, even when using the same or even fewer total participants. A comprehensive analysis of staircase design including optimal cluster proportion to the sequences, appropriate cluster sizes, and Bayesian design strategies should be thoroughly explored.

8. Discussion and conclusion

In this review we have traced the evolution of stepped wedge designs (SWDs) from the foundational Hussey–Hughes random-intercept model through modern Bayesian and "staircase" alternatives. A recurring theme is the trade–off between analytical simplicity and realistic correlation structure. Early formulas for power and sample size assume constant ICCs, cross-sectional sampling, and equal cluster sizes; these yield closed-form design effects

but can mislead when within- and between-period correlations differ or clusters vary in size. Extensions to cohort and open-cohort SWDs introduced additional ICC parameters and decay models, but at the cost of analytical intractability. Simulation-based approaches remedy this at the expense of computational burden and reliance on assumed nuisance parameters. Our survey of optimal-allocation methods highlights how design efficiency depends critically on the allocation of clusters to sequences. Bayesian D-optimal frameworks then add robustness by placing priors on ICC or time-effect parameters, reducing required sample size when external information is reliable but risking bias under prior misspecification. Lastly, the staircase design represents a pragmatic compromise: by sampling only around each cluster's switch point, it retains most information on treatment contrasts while cutting data-collection burden. Across a broad range of ICC scenarios, basic staircase trials can even outperform full SWDs in power per participant.

Despite these advances, several gaps remain. First, most methods target continuous outcomes; extensions to binary, count or time–to–event endpoints require further development. Second, while correlation-decay models are conceptually appealing, real world validation via intensive pilot data or retrospective re-analysis of completed SWDs remains scarce. Third, the increasing complexity of hybrid designs (*e.g.* unequal cluster sizes, open cohorts, Bayesian priors) calls for user friendly software that integrates power, sample size, and optimal allocation routines under a unified interface. Finally, practical considerations such as staggered enrollment logistics, missing data, and secular trends–deserve more attention in design-stage simulations.

In sum, the stepped wedge framework has matured from simple cross sectional formulas to a rich design space encompassing complex correlation structures, Bayesian robustness, and lean staircase variants. The choice among these should be driven by the substantive context–outcome type, anticipated ICC patterns, logistical constraints and cluster sizes. In this review paper, we not only mention a few relevant works in various field of SW design but also explain fundamental terminologies related to this design in a concise manner, aiming to assist readers who are encountering these concepts for the very first time. We hope this introductory yet informative overview provides a solid foundation for further exploration into the field of SW design.

Acknowledgements

The work of Soumadeb Pain is funded through an IIT Kanpur assistantship, funded by the Ministry of Education(MoE), Govt. of India. The work of Satya Prakash Singh is funded by Science and Engineering Research Board, Grant/Award Number: MTR/2022/000627. We are indeed grateful to the Editors for their guidance and counsel. We also sincerely appreciate the reviewer's valuable comments, insightful suggestions, and the generous inclusion of numerous useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., and Omar, R. Z. (2015). Sample size calculation for a stepped wedge trial. *Trials*, **16**.

Beard, E., Lewis, J. J., Copas, A., Davey, C., Osrin, D., and Baio, G. e. a. (2015). Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials*, **16**.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, **6**.

Copas, A. J., Lewis, J. J., Thompson, J. A., Davey, C., Baio, G., and Hargreaves, J. R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials*, **16**, 1–12.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomised Trials in Health Research*. Arnold, London.

Etfer, L., Wason, J., and Grayling, M. J. (2024). Optimal Bayesian stepped-wedge cluster randomised trial designs for binary outcome data. *arXiv preprint arXiv:2402.09938*, **abs/2402.09938**, 1–15.

Ford, W. P. and Westgate, P. M. (2020). Maintaining the validity of inference in small-sample stepped wedge cluster randomized trials with binary outcomes when using generalized estimating equations. *Statistics in Medicine*, **39**, 2779–2792.

Gasparini, A., Crowther, M. J., Hoogendijk, E. O., Li, F., and Harhay, M. O. (2025). Analysis of cohort stepped wedge cluster-randomized trials with nonignorable dropout via joint modeling. *Statistics in Medicine*, **44**, e10347.

Girling, A. J. (2018). Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Statistics in Medicine*, **37**, 4652–4664.

Girling, A. J. and Hemming, K. (2016). Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine*, **35**, 2149–2166.

Grantham, K. L., Forbes, A. B., Hooper, R., and Kasza, J. (2024). The staircase cluster randomised trial design: a pragmatic alternative to the stepped wedge. *Statistical Methods in Medical Research*, **33**, 24–41.

Grantham, K. L., Forbes, A. B., Hooper, R., and Kasza, J. (2025). The relative efficiency of staircase and stepped wedge cluster randomised trial designs. *Statistical Methods in Medical Research*, **34**, 701–716.

Grantham, K. L., Kasza, J., Heritier, S., Hemming, K., and Forbes, A. B. (2019). Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in Medicine*, **38**, 1918–1934.

Hall, A. J., Inskip, H. M., and Loik, F. e. a. (1987). The gambia hepatitis intervention study. *Cancer Research*, **47**, 5782–5787.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., and Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, **350**.

Hemming, K. and Taljaard, M. (2016). Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology*, **69**, 137–46.

Hemming, K. and Taljaard, M. (2020). Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *International Journal of Epidemiology*, **49**, 1043–1052.

Hemming, K., Taljaard, M., and Forbes, A. (2018). Modeling clustering and treatment effect heterogeneity in parallel and stepped-wedge cluster randomized trials. *Statistics in Medicine*, **37**, 883–898.

Hooper, R. and Copas, A. (2019). Stepped wedge trials with continuous recruitment require new ways of thinking. *Journal of Clinical Epidemiology*, **116**, 161–166.

Hooper, R., Kasza, J., and Forbes, A. (2020). The hunt for efficient, incomplete designs for stepped wedge trials with continuous recruitment and continuous outcome measures. *BMC Medical Research Methodology*, **20**, 1–9.

Hooper, R., Teerenstra, S., de Hoop, E., and Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, **35**, 4718–4728.

Hughes, J. P., Granston, T. S., and Heagerty, P. J. (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, **45**, 55–60.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, **28**, 182–191.

Kasza, J. and Forbes, A. B. (2019). Inference for the treatment effect in multiple-period cluster randomised trials when random effect correlation structure is misspecified. *Statistical Methods in Medical Research*, **28**, 3112–3122.

Kasza, J., Hemming, K., Hooper, R., Matthews, J., and Forbes, A. (2019). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, **28**, 703–716.

Kasza, J., Hooper, R., Copas, A., and Forbes, A. B. (2020). Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in Medicine*, **39**, 1871–1883.

Kenny, A., Voldal, E. C., Xia, F., Heagerty, P. J., and Hughes, J. P. (2022). Analysis of stepped wedge cluster randomized trials in the presence of a time-varying treatment effect. *Statistics in Medicine*, **41**, 4311–4339.

Kristunas, C. A., Smith, K. L., and Gray, L. J. (2017). An imbalance in cluster sizes does not lead to notable loss of power in cross-sectional, stepped-wedge cluster randomised trials with a continuous outcome. *Trials*, **18**, 1–11.

Lawrie, J., Carlin, J. B., and Forbes, A. B. (2015). Optimal stepped wedge designs. *Statistics and Probability Letters*, **99**, 210–214.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, **19**, 219–238.

Li, F. (2020). Design and analysis considerations for cohort stepped wedge cluster randomized trials with a decay correlation structure. *Statistics in Medicine*, **39**, 438–455.

Li, F., Turner, E. L., and Preisser, J. S. (2018a). Optimal allocation of clusters in cohort stepped wedge designs. *Statistics and Probability Letters*, **137**, 257–263.

Li, F., Turner, E. L., and Preisser, J. S. (2018b). Sample size determination for gee analyses of stepped wedge cluster randomized trials. *Biometrics*, **74**, 1450–1458.

Li, F. and Wang, R. (2022). Stepped wedge cluster randomized trials: a methodological overview. *World Neurosurgery*, **161**, 323–330.

Li, F., Yu, H., Rathouz, P. J., Turner, E. L., and Preisser, J. S. (2022). Marginal modeling of cluster-period means and intraclass correlations in stepped wedge designs with binary outcomes. *Biostatistics*, **23**, 772–788.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Martin, J. T., Hemming, K., and Girling, A. (2019). The impact of varying cluster size in cross–sectional stepped-wedge cluster randomised trials. *BMC Medical Research Methodology*, **19**.

Matthews, J. N. S. (2020). Highly efficient stepped wedge designs for clusters of unequal size. *Biometrics*, **76**, 1167–1176.

Mdege, N. D., Man, M.-S., Taylor, C. A., and Torgerson, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, **64**, 936–948.

Ouyang, Y., Karim, M. E., Gustafson, P., Field, T. S., and Wong, H. (2020). Explaining the variation in the attained power of a stepped-wedge trial with unequal cluster sizes. *BMC Medical Research Methodology*, **20**.

Singh, S. P. (2024). Bayesian optimal stepped wedge design. *Biometrical Journal*, **66**, 2300168.

Thompson, J., Hemming, K., Forbes, A., Fielding, K., and Hayes, R. (2021). Comparison of small sample standard error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: a simulation study. *Statistical Methods in Medical Research*, **30**, 425–439.

Thompson, J. A., Fielding, K., Hargreaves, J., and Copas, A. (2017). The optimal design of stepped wedge trials with equal allocation to sequences and a comparison to other trial designs. *Clinical Trials*, **14**, 639–647.

Turner, E. L., Li, F., Gallis, J. A., Prague, M., and Murray, D. M. (2017a). Review of recent methodological developments in group-randomized trials: part 1—design. *American Journal of Public Health*, **107**, 907–915.

Turner, E. L., Prague, M., Gallis, J. A., Li, F., and Murray, D. M. (2017b). Review of recent methodological developments in group-randomized trials: part 2—analysis. *American Journal of Public Health*, **107**, 1078–1086.

Varghese, E., Briola, A., Kennel, T., Pooley, A., and Parker, R. A. (2025). A systematic review of stepped wedge cluster randomized trials in high impact journals: assessing the design, rationale, and analysis. *Journal of Clinical Epidemiology*, **178**, 111622.

Wang, J., Cao, J., Zhang, S., and Ahn, C. (2021). Sample size and power analysis for stepped wedge cluster randomised trials with binary outcomes. *Statistical Theory and Related Fields*, **5**, 162–169.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., and Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, **66**, 752–758.

Xia, F., Hughes, J. P., Voldal, E. C., and Heagerty, P. J. (2021). Power and sample size calculation for stepped-wedge designs with discrete outcomes. *Trials*, **22**, 1–10.

Zhan, D., Ouyang, Y., Xu, L., and Wong, H. (2021a). Improving efficiency in the stepped–wedge trial design via Bayesian modeling with an informative prior for the time effects. *Clinical Trials*, **18**, 295–302.

Zhan, D., Xu, L., Ouyang, Y., Sawatzky, R., and Wong, H. (2021b). Methods for dealing with unequal cluster sizes in cluster randomized trials: A scoping review. *PLoS ONE*, **16**, e0255389.

Zhan, Z., de Bock, G. H., and van den Heuvel, E. R. (2018). Optimal unidirectional switch designs. *Statistics in Medicine*, **37**, 3573–3588.

Zhou, X., Liao, X., Kunz, L. M., Normand, S.-L. T., Wang, M., and Spiegelman, D. (2020). A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics*, **21**, 102–121.