Duke GLOBAL HEALTH INSTITUTE

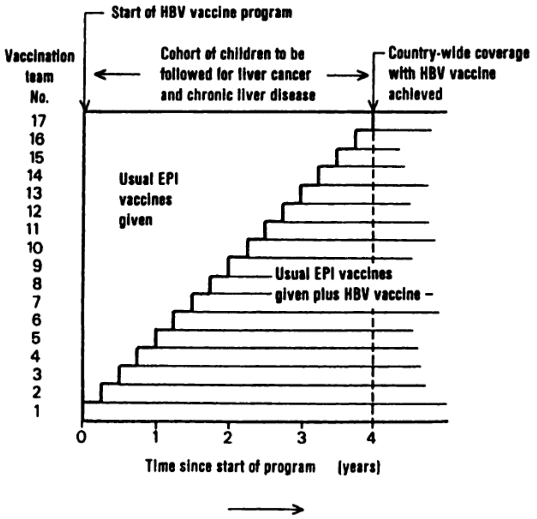
Research Design & Analysis Core

Core Guide: Stepped Wedge Cluster Randomized Designs

Part of a series addressing common issues in statistical and epidemiological design and analysis

Background and overview

Definition: A SW-CRT, is a type of crossover cluster randomized trial in which all clusters start in the control condition and are randomized to the time point when the intervention with be implemented. A key feature of a SW-CRT is that every cluster eventually receives the intervention. Copas et al. (2015) define a SW-CRT as "**a trial in which clusters receive the intervention at different time points, the order in which they receive it is randomized, and data are collected from clusters over time**."


Start of HBV vaccine program
Vaccination team No.
Cohort of children to be followed for liver cancer and chronic liver disease
Country-wide coverage with HBV vaccine achieved
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Usual EPI vaccines given
Usual EPI vaccines given plus HBV vaccine
0
1
2
3
4
Time since start of program (years)

g. 2. Phased introduction of hepatitis B vaccination in The Gambia.

Background and Example: According to Rhoda, Murray, Andridge, Pennell, and Hade (2011), the name "stepped wedge originated with the Gambia Hepatitis Study and refers to the wedge shape of the intervention timeline across groups" (Gambia Hepatitis Study Group, 1987). That is, the stepped wedge design can be visualized as a wedge shape with various steps indicating the start of implementation of the intervention for a particular cluster. For example, Figure 1 is extracted from The Gambia Hepatitis Study paper and shows a staggered implementation of the new treatment (usual [EPI] vaccines plus HBV vaccines). The goal of that study was to estimate the effectiveness of adding a Hepatitis B Virus (HBV) vaccine to the existing schedule of usual (EPI) vaccines that were given in childhood. To accomplish this, the 17 teams administering usual (EPI) vaccines were randomly assigned starting times for adding HBV vaccines to their list.

The Gambia Hepatitis Study Group provided the following reasons for their choice of the stepped wedge design:

- "the expense of the vaccine and its limited availability prohibiting immediate universal HBV vaccination;
- The desirability of having comparison groups available from the same time period;

- The severe logistic difficulties that would have been encountered with randomization at the individual level…;
- The hope that HBV vaccine would be widely available at the end of the study….”

***Figure 1.* Extract from Gambia Hepatitis Study (1987) showing SW-CRT design**

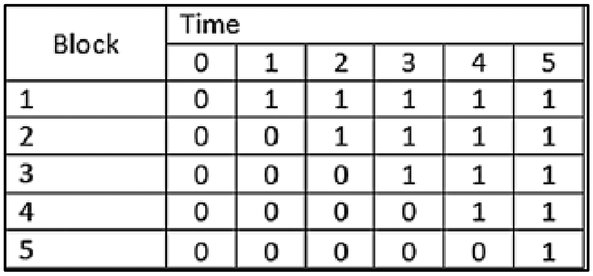
In this example and as noted by the authors, it was natural for the intervention to be rolled out to clusters of children rather than to use an individually-randomized design. The stepped wedge design can also be applied in the individually-randomized trial setting and when randomization is not used. Because randomization is most commonly used to assign the implementation start of the intervention and because the unit of randomization is typically a cluster rather than an individual, we focus on the stepped wedge cluster randomized trial (SW-CRT) design in the current guide.

Alternative names: Stepped wedge designs have been given various names in the applied literature. The following is a list of some of the more common names:

- Delayed intervention
- Delayed treatment
- Waiting list
- Phased implementation
- Staggered implementation
- Crossover

However, crossover trials are a more general, broader class of trials because they include designs where clusters cross back over into a control condition after receiving the intervention.

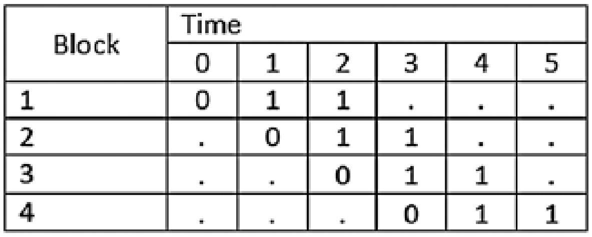
Types: In a **complete stepped wedge** design, all clusters are measured at every time point, regardless of when they receive the intervention. This is illustrated in Figure 2. Although the intervention is not implemented until time period 5 for all the clusters that appear in block 5, outcomes and predictors are still measured at all preceding time points.



Block	Time					
	0	1	2	3	4	5
1	0	1	1	1	1	1
2	0	0	1	1	1	1
3	0	0	0	1	1	1
4	0	0	0	0	1	1
5	0	0	0	0	0	1

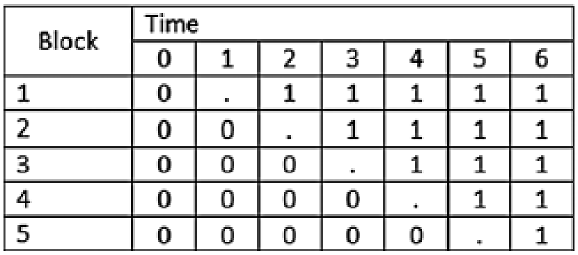
Figure 2. From Hemming, Lilford, and Girling (2015).

In an **incomplete stepped wedge** design, not every cluster is measured at every time point. An example from Hemming, Lilford, et al. (2015) is given in Figure 3. In this case, the clusters in block 4 are first measured in time period 3, even though other blocks are measured at previous times. This type of design could arise, for example, if resource constraints limit the number of clusters that could be measured at each time point. Another illustration, given in Figure 4, shows an incomplete design with an implementation period (denoted by ".") during which outcomes are not measured.



Block	Time					
	0	1	2	3	4	5
1	0	1	1	.	.	.
2	.	0	1	1	.	.
3	.	.	0	1	1	.
4	.	.	.	0	1	1

Figure 3.



Block	Time						
	0	1	2	3	4	5	6
1	0	.	1	1	1	1	1
2	0	0	.	1	1	1	1
3	0	0	0	.	1	1	1
4	0	0	0	0	.	1	1
5	0	0	0	0	0	.	1

Figure 4.

Rationale and concerns: Common reasons given for implementing a SW-CRT instead of a classic parallel CRT include logistical (e.g., do not have enough resources to deliver the intervention all at once) and ethical (e.g., all clusters should receive intervention eventually). It is now widely accepted that the primary rationale for the SW-CRT design choice should be logistical because, if there are ethical concerns, a parallel-arm CRT could be implemented with rollout to control arm clusters at the end of the evaluation period if there is evidence of effectiveness of the intervention. Nevertheless, in a recent systematic review by Mdege, Man, Taylor, and Torgerson (2011), the authors note that most of the 25 studies they included in their review cited ethical reasons, stating that denying the intervention to any group would be considered unethical.

An interesting summary of a half-day conference on stepped wedge trials (SW-CRTs) held at the London School of Hygiene and Tropical Medicine can be found at this link. The author reviews some of the discussions which took place at the conference concerning the strengths and weaknesses of SW-CRTs.

Since SW-CRTs generally take longer to implement than a traditional parallel CRT, stepped wedge trials should generally be used in studies for logistical reasons, such as resource constraints. A good example of this is an intervention delivered in small groups in a resouce-limited setting. Because of the nature of the intervention, it is only possible to implement it with two groups of 8 at one time. In order to obtain 96 participants and have more power to detect an intervention effect, a SW-CRT design is necessary. However, if this resource constraint were not in place, then it would make sense to conduct a parallel CRT with all clusters at once. Then if it is considered unethical to withhold the intervention, the clusters assigned to control could be given the intervention at the end of the trial. If it is considered unethical to even delay the treatment, then some type of "enhanced usual care" treatment should be given to the controls while they wait.

Another consideration, in addition to that mentioned in the previous paragraph, is the possibility for increased statistical efficiency with SW-CRT designs versus parallel-arm CRTs. This will be discussed more in the **sample size calculations** section below.

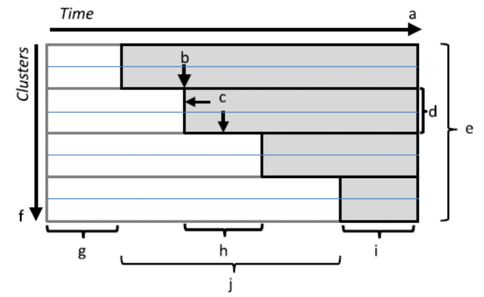
Design considerations

Copas et al. (2015) describe many of the design features of SW-CRTs. They identify several key aspects of the allocation strategy:

- Number of clusters per group
- Number of groups
- The length of time between successive crossover points ("step length")
- Total number of clusters
- Trial duration

The authors provide a helpful diagram, which is reproduced in Figure 5. They then describe three main types of designs for participant recruitment, and provide case studies.

- **Continuous recruitment with short exposure.** As the trial begins, very few individuals participate, but over time, more become eligible and are then exposed to intervention for a short time, after which the outcome measure is typically measured.
- **Closed cohort.** "All participants are identified at the [outset] of the trial and participate from start till end," and typically participants do not change clusters during the trial. "Repeated measurements are typically taken from the same individuals to assess change and its relation to exposure" (Copas et al., 2015).
- **Open cohort.** "A substantial number of individuals are identified and participate from the start, but some may leave during the trial and others may become eligible and be exposed for some time. A minority of individuals may also change between trial clusters" (Copas et al., 2015).a



a. Duration of trial
b. Crossover point
c. (2nd) step
d. Number of clusters in (2nd) group
e. Number of groups
f. Total number of clusters
g. Pre-rollout period
h. Time between (2nd and 3rd) crossover points
i. Post-rollout period
j. Rollout period

Fig. 1 Characteristics and terminology of stepped-wedge cluster randomised controlled trials, where shaded areas indicate intervention exposure and unshaded areas indicate control exposure

Figure 5.

Much of the methods literature has focused on a small range of SW-CRT designs where data are collected from individuals at discrete time points, and individuals contribute one measurement during the study (Hemming, Lilford, et al., 2015; Hussey & Hughes, 2007). In the literature, these are often called **cross-sectional** designs. However, Copas et al. (2015) found that much of the published literature follow a cohort design. The paper discusses various designs and the issues surrounding them.

Girling and Hemming (2016) discuss statistical efficiency and design in the context of linear mixed effects models. They examine both cross-sectional and longitudinal closed cohort designs, adding cluster-specific time components and subject-level components in the longitudinal design. In their paper, statistical efficiency is defined in a formula which uses the cluster-mean correlation (CMC), which is defined as the proportion of the variance of a cluster-mean "that comes from random effects that are independent of time." They conclude that in large studies "the best design [in terms of efficiency] is a hybrid mixture of parallel and stepped wedge components", although the authors provide a simple search algorithm for determining optimal designs when results pertaining to large studies are inapplicable. In addition, using the CMC, researchers now have a relatively simple way to compute design effects, which are discussed in the next section.

Sample size calculations: In the traditional parallel-arm CRT, required sample size to achieve a desired level of power is often computed by first calculating sample size needed for an individually randomized trial, and then adjusting the sample size by a variance inflation factor, usually expressed as a design effect (Baio et al., 2015). The literature is rich on approaches for sample size calculation for classical CRTs; however, the literature is much less developed for SW-CRTs.

In addition to the standard features of CRTs, such as number of clusters and cluster size, SW-CRT sample size calculations must also take into account "the number of crossover points; the number of clusters switching intervention arm at each time point; possible time and/or lag effect..."; and "whether the data are collected for a SW-CRT in a cross-sectional manner or they are repeated measurements on the same individuals" (Baio et al., 2015).

The seminal paper on sample size calculation for SW-CRTs was authored by Hussey and Hughes (2007), and they've recently created an R package for implementing these calculations (`swCRTdesign`). Although they provide an analytic formula for computing sample size, it is of limited application. As Hemming, Lilford, et al. (2015) point out, the SW-CRT designs considered by Hussey and Hughes (2007) are for situations "where data are obtained or collected for the analysis at each and every step," which is called a **complete design** (see figure 2). In addition, Hussey and Hughes (2007) only consider designs in which (as mentioned previously in this guide) data are collected from individuals at discrete time points, and individuals contribute one measurement during the study (cross-sectional design). For such specific designs, the formula provided by Hussey and Hughes (2007) may be useful.

Hemming, Lilford, et al. (2015) and Hemming and Taljaard (2016) expand on the work of Hussey and Hughes (2007) by providing sample size calculation strategies for designs other than the complete design, although *their formulae are still only applicable to cross-sectional designs*. Their formulae also allow for designs with multiple levels of clustering. In their approach, the researcher provides the information about the variance-covariance matrix of cell means, which allows flexibility in specifying incomplete designs. In order to implement these calculations, Hemming and Girling (2014) provide a Stata program for computing sample size, which allows for continuous, binary, and count outcomes.

In Baio et al. (2015) the authors cite the review by Beard et al. (2015) showing that, of the 38 SW-CRT studies identified in the review, 10 used formulae for cluster RCTs and 8 did not report sample size calculation. Of those accounting for the SW-CRT design, the researchers used various methods for calculating sample size, including simulations (Hussey & Hughes, 2007; Moulton et al., 2007; Woertman et al., 2013).

Baio et al. (2015) is an excellent review of the SW-CRT sample size literature up to the year 2015. All of the analytic methods they review are based on cross-sectional designs. These methods will tend to overestimate the required sample size in the more common cohort designs (assuming some level of positive correlation between measurements on the same individual over time). Baio et al. (2015) suggest researchers use a simulation-based approach for sample size calculations in SW-CRTs, given the complexity and variety of SW-CRT designs. As of August 17, 2016, Baio has provided an R package (currently in Beta level of development) for simulating power for both cross-sectional and cohort designs, as well as for applying analytic formulae, such as the formula found in Hussey and Hughes (2007), for simple designs (Baio).

Generally, the following increase power in a SW-CRT design: a lower ICC, more participants, and an increased number of steps (cross-over points). Because of the confounding effect of calendar time, the standard parallel-arm CRT design effect is not applicable (Hemming, Haines, Chilton, Girling, & Lilford, 2015).

It should be noted that in some situations, SW-CRT designs can provide increased statistical efficiency over classical parallel-arm CRT designs. Baio et al. (2015) conclude that, under the assumptions of a cross-sectional design where each participant provides one measurement, their results suggest that "a SW-CRT is more efficient unless the ICC is rather low, for example, much less than 0.1. In other words, given cross-sectional data and the same number of participants measured per cluster, the SWT may often be a more efficient trial design and so will require fewer clusters." Hemming, Haines, et al. (2015) also show this in their article. An SW-CRT may often be more efficient and require fewer clusters. Cost and design should be taken into account.

Analysis considerations

In the SW-CRT design, "the distribution of results across unexposed observation periods is compared with that across the exposed observation periods". This means that both within-cluster and between-cluster comparisons can be used to estimate the intervention effect, in contrast to the parallel-arm CRT design in which only between-cluster information is available because the cluster is observed either under control or under intervention (Hemming, Haines, et al., 2015).

Analysis will depend on the specific design used. Like classic parallel CRTs, SW-CRTs must take into account various levels of correlation. These include correlation within individuals over time and between individuals in each level of clustering. SW-CRTs pose additional challenges beyond simply accounting for the correlation structure. Particularly, as Davey et al. (2015) point out, "in SW-CRTs the effect estimate is potentially confounded by secular changes in the outcome." For example, the outcome measured in an SW-CRT rolling out a vaccine intervention (e.g., The Gambia Hepatitis Intervention Study) could be affected by seasonal variation in infection rates. Because of the way the intervention is rolled out to clusters in the SW-CRT, control (unexposed) observations will be, on average, from an earlier point in time than treatment (exposed) observation. As a consequence, time is a potential confounder and it should be adjusted for in the analysis. One of the best references on formulating a statistical model in stepped wedge designs is Hemming, Lilford, et al. (2015), in which the authors provide notation and details on the analysis of stepped wedge designs.

Davey et al. (2015) discuss two approaches to analyzing data from a SW-CRT. Here, we will briefly discuss the more widely used approach. In this method, the statistician "explicitly takes into account secular trends by producing an intervention effect adjusted for time trends, which are also estimated." "Time trends are commonly entered into the model as fixed effects, often as factors simply reflecting the periods between crossover points, with the assumption that the trend is the same in all clusters."

To take into account clustering in the data (and possibly the longitudinal nature of the data), SW-CRTs will generally be analyzed using either generalized linear mixed models (GLMMs) or generalized estimating equations (GEE). Choice of type of model should be based on underlying assumptions and design considerations.

Some other analysis issues to consider in SW-CRT designs is a lag in the intervention effect and loss of fidelity in the intervention over time (Hughes, Granston, & Heagerty, 2015).

Reporting: Hemming, Haines, et al. (2015) provide a table with suggestions to modify the Consort 2010 extension for reporting SW-CRTs. They include such things as design rationale (why choose a SW-CRT), length of steps, sample size justification, and some analysis considerations. The fact that the clusters are randomized to start time of the intervention, rather than to intervention and control as in CRTs, may make reporting more difficult. "The large number of allocation groups and the crossover from control to intervention allocation can make it less straightforward to present a participant flow (that is, CONSORT diagram) for SW-CRTs relative to CRTs" (Davey et al., 2015). It is also more difficult to assess and report balance by group, since there are a large number of groups to which a small number of clusters per group are allocated. In addition, while researchers may want to report balance by treatment condition, balance usually also depends on secular trends in the outcome. A good example of reporting is for the Expedited Partner Therapy trial conducted in 23 health clusters in Washington State and which was the motivating example for the original Hussey and Hughes (2007) paper (Golden et al., 2015).

Davey et al. (2015) found significant heterogeneity in the reporting of SW-CRTs, which may be because of the novelty of the design, as well as the lack of a reporting guideline (Consort statement). Martin, Taljaard, Girling, and Hemming (2016) also discovered that reporting in SW-CRTs is sub-optimal. Until a more detailed Consort statement is produced, researchers should use the guidelines for CRTs, with the additions and modifications proposed by Hemming, Haines, et al. (2015).

Prepared by: John Gallis, MSc
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Version: 1.4, last updated August 31, 2017

References

Baio, G. SWSamp: Simulation-based sample size calculations for a Stepped Wedge Trial (and more).

Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., & Omar, R. Z. (2015). Sample size calculation for a stepped wedge trial. *Trials, 16*(1), 354.

Beard, E., Lewis, J. J., Copas, A., Davey, C., Osrin, D., Baio, G., . . . Ononge, S. (2015). Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials, 16*(1), 1-14.

Copas, A. J., Lewis, J. J., Thompson, J. A., Davey, C., Baio, G., & Hargreaves, J. R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials, 16*(1), 352.

Davey, C., Hargreaves, J., Thompson, J. A., Copas, A. J., Beard, E., Lewis, J. J., & Fielding, K. L. (2015). Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials, 16*(1), 358.

Gambia Hepatitis Study Group. (1987). The Gambia hepatitis intervention study. *Cancer Research, 47*(21), 5782-5787.

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

Golden, M. R., Kerani, R. P., Stenger, M., Hughes, J. P., Aubin, M., Malinski, C., & Holmes, K. K. (2015). 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 Medicine, 12*(1), e1001777.

Hemming, K., & Girling, A. (2014). A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal, 14*, 363-380.

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

Hemming, K., Lilford, R., & Girling, A. J. (2015). Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Statistics in Medicine, 34*(2), 181-196.

Hemming, K., & Taljaard, M. (2016). Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology, 69*, 137-146.

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

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*(2), 182-191.

Martin, J., Taljaard, M., Girling, A., & Hemming, K. (2016). Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ open, 6*(2), e010166.

Mdege, N. D., Man, M.-S., Taylor, C. A., & 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*(9), 936-948.

Moulton, L. H., Golub, J. E., Durovni, B., Cavalcante, S. C., Pacheco, A. G., Saraceni, V., . . . Chaisson, R. E. (2007). Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials, 4*(2), 190-199.

Rhoda, D. A., Murray, D. M., Andridge, R. R., Pennell, M. L., & Hade, E. M. (2011). Studies with staggered starts: multiple baseline designs and group-randomized trials. *American Journal of Public Health, 101*(11), 2164-2169.

Woertman, W., de Hoop, E., Moerbeek, M., Zuidema, S. U., Gerritsen, D. L., & Teerenstra, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology, 66*(7), 752-758.