OPEN ACCESS

Cutting edge or blunt instrument: how to decide if a stepped wedge design is right for you

Richard Hooper, Sandra M Eldridge

Institute of Population Health Sciences, Queen Mary University of London, London, UK

Correspondence to
Prof Richard Hooper, Queen Mary University of London, London E1 2AB, UK;
r.l.hooper@qmul.ac.uk

Received 25 May 2020
Accepted 26 May 2020
Published Online First 16 June 2020

To cite: Hooper R, Eldridge SM. *BMJ Qual Saf* 2021;**30**:245–250.

INTRODUCTION

The last 10 years have seen an extraordinary surge of interest in 'stepped wedge' designs for evaluating interventions to improve health and social care. Reviews of published trials and registered protocols have shown an exponential increase in the number of trials citing a stepped wedge approach.[1–6] A growing body of work on methods for the design, conduct and analysis of stepped wedge trials has emerged, building on seminal work by Hussey and Hughes in 2007.[7] The Consolidated Standards of Reporting Trials reporting guidelines for stepped wedge cluster randomised trials are now available, making it easier for investigators to appraise evidence and plan their own evaluations.[8]

But published examples of stepped wedge evaluations in quality improvement illustrate some of the practical challenges. On the one hand, limited research resources may force investigators to stagger implementation at different sites[9]; on the other hand, persuading sites to follow a precise, predetermined schedule for implementation may be hard.[10] In fact, investigators who plan a stepped wedge trial must balance a number of logistical, ethical and methodological issues.[11 12] In this article, we focus predominantly on the design of such evaluations, and encourage a questioning approach. We take a 'trial' to mean a study involving the prospective, experimental allocation of interventions,[13] but more particularly we focus on studies where those allocations are randomised. We start with the question of what is meant by a stepped wedge trial.

WHAT IS A STEPPED WEDGE CLUSTER RANDOMISED TRIAL?

The vast majority of stepped wedge trials are cluster randomised, and when people refer to stepped wedge designs this is usually what they have in mind. A cluster randomised trial is a trial in which all the participants at the same site or 'cluster' are allocated to the same intervention.[14] Stepped wedge cluster randomised trials are run over an extended interval of time, allowing clusters to cross over from a routine care or 'control' condition to an experimental intervention condition *during the trial*.[15] This means that as well as comparing clusters concurrently under different conditions, you can compare participants in the same cluster before and after the introduction of the intervention. In the most common scheme, all clusters begin in the control condition, finish in the intervention condition and cross over at evenly spaced intervals. This mimics many natural (non-experimental) implementation processes, and stepped wedge trials are widely seen as useful for evaluating policy changes and other interventions that were due to be 'rolled out anyway'.[2]

Exactly what it means for the timescale to be 'extended' will depend on the trial. Stepped wedge trials come in many and varied forms.[16] One approach is to recruit all the participants at the start of the trial, and to follow them prospectively as a cohort. For instance, an evaluation of an emergency admission risk prediction tool in primary care, randomised by general practice, followed a single cohort of patients registered with participating practices at the start of the trial period, who were tracked throughout the trial. Each month more of the practices switched over to using the tool, according to a randomised timetable.[17]

The same study also took a series of cross-sectional samples from the larger cohort of patients (not necessarily the same patients each time) to assess quality of life and satisfaction.[17] This

Box 1 Practical constraints on the design of a longitudinal cluster randomised trial

- Are there limits on the time available to complete the evaluation, on the number of clusters, or on the number of participants (or the rate at which you can recruit participants) at each cluster? These constraints put limits on the overall scale of the evaluation, or force trade-offs between different design characteristics.
- How will participants and their data be sampled in your study: as a series of cross-sectional surveys, as a continuous stream of incident cases, as a cohort followed over time, or some other way? Does the timescale divide into cycles, seasons or milestones that influence how you will sample participants and data?
- Is there a limit on how many clusters can implement the intervention at the same time in the evaluation? If this is constrained by research resources (eg, if there are only enough trained research staff to implement the intervention one cluster at a time) then implementation *must* be staggered in some way.
- If implementation is to be staggered, is there a minimum 'step length'? If the same team delivers the intervention in different clusters at different steps, then bear in mind it may take some time to get the intervention fully operational at a site, and the team will also need time to relocate from one cluster to the next.

repeated cross-sectional approach offers another way of conducting a stepped wedge trial. Extending the timescale in this case simply means scheduling more cross-sectional surveys, with clusters (practices in this example) crossing from the control to the intervention between successive surveys.

A more common approach is to recruit eligible participants as they present at clusters in a continuous stream.[18] In this case, a longer recruitment period leads to more participants and more time to cross clusters over. For instance, in a stepped wedge evaluation of an intrapartum emergencies training package, eligible women were included as and when they gave birth at 12 maternity units (clusters) in Scotland.[10] The investigators anticipated that for every 6 months they extended recruitment they could identify, on average, 1200 more births per cluster (maternity unit). A different batch of maternity units was crossed over to the intervention every 6 months.

WHEN MIGHT I CONSIDER DOING A STEPPED WEDGE TRIAL?

Research designs are shaped as much by practical constraints as by abstract schemes, and it is always a good idea to start with the constraints and work towards a design, rather than start with a design and try to fit it to constraints. These constraints will be unique to each research context, and box 1 lists some areas to think about. Still, there are some common features of settings where a stepped wedge trial might be considered as a possible design, and we now review these.

Stepped wedge trials are suited to situations where, while it might be easy enough to introduce the experimental intervention to a cluster, it is much harder (practically or politically) to take it away again. These are interventions that change practice or are difficult to unlearn, or that policy has decreed will be rolled out anyway. This restriction is sometimes referred to as one-way crossover. (There are certainly interventions that can be crossed both ways, from control to intervention and back again, but in this case a design with *two-way* crossover—distinct from a stepped wedge—is recommended: we leave further discussion of these cluster randomised cross-over trials to others.)[19 20]

Stepped wedge designs also implicitly require that all of the clusters that will participate in the trial are ready to start (to be randomised and commence data collection) at the same calendar date—in other words that there is no long, drawn-out period of recruitment of sites. Studies where site recruitment will be a drawn-out process must follow an alternative strategy where each cluster is randomised as and when it is recruited, either to the control or to the intervention—just as you would randomise individuals in the simplest design for an individually randomised trial.

Remember, also, that one defining feature of a stepped wedge trial is that it runs over an extended time period. One of the most important questions to ask is whether this is necessary at all. In research on health services and quality improvement, marshalling good evidence *quickly* is likely to trump most other considerations of research design. So, if you can gather all the evidence you need *without* having to schedule repeated visits to your sites over months or years, or stagger the implementation of the intervention at different sites, then this is what you should do. We reflect further on some of these issues below.

The motivation for conducting a stepped wedge trial that is most commonly cited is also the most questionable: that a stepped wedge design is necessary when you want everyone to have the opportunity to access the intervention. This is often portrayed as an incentive for sites to participate, or as an ethical obligation, or as a justification based on a concern that sites might seek the intervention for themselves outside of the trial protocol. We will square up to the logic of this argument in the next section.

A much more pertinent question to ask than 'should I give every site the intervention?' is 'how long can I reasonably ask any site to wait for it?' This will help you understand how much time you have to conduct a truly randomised evaluation. If you believe, incidentally, that you have an ethical obligation to give everyone the intervention immediately, and if you can,

Box 2 How the figures for statistical power in figure 1 were calculated

Sample size calculations for trials usually determine the number of participants needed to achieve given statistical power,[28] but here we illustrate the power achieved with different design choices assuming that the number of clusters (maternity units) is fixed at 10. Four women are recruited every month at each cluster. Cluster randomised trials generally have less power than individually randomised trials because of the similarity of the outcomes of individuals who belong to the same cluster: this is quantified by the intracluster correlation coefficient (ICC).[36] Here we assume that the ICC for any two women attending the same maternity unit is 0.01. The other consideration crucial to the power is the minimal clinically important intervention effect we would like to have power to demonstrate.[37] For illustration, we assume we want power to demonstrate a mean difference of 0.4 times the SD in our primary outcome measure. We have used methods for calculating power that are described elsewhere.[36 38–40] These calculations assume we are adjusting for possible changes in outcomes over time. All statements of power are at the 5% significance level.

then a stepped wedge trial is *not* appropriate (nor is any kind of trial). It would be as unethical, in this case, to randomise some sites to wait for the intervention as it would be to randomise half to the intervention and half to control.[12]

DO I *NEED* TO USE A STEPPED WEDGE DESIGN?

So, what if we have an intervention that can only be crossed in one direction, and we have a number of clusters that are ready to be randomised at the same time to a trial conducted over an extended period of time. How do we arrive at a stepped wedge as our design choice rather than any alternative?

Suppose we want to design a trial in a maternity unit setting, recruiting women with suspected pre-eclampsia, and randomised by maternity unit. Suppose we have identified 10 maternity units willing to take part, and we are not hopeful of finding any more. For this example, we will divide the timetable for the study into whole months for convenience and assume that in each unit four women are recruited every month. Here we explore the statistical power—the likelihood of finding evidence for an important effect—of different designs. More details on the assumptions behind our power calculations are given in box 2.

First, a sense-check: do we really need to extend the timescale of our trial? What if we recruited women over a single month, with half the maternity units allocated to the intervention condition and half to the control (20 women in each condition)? This design is shown schematically in figure 1A. The power is 24%—not great, as we usually aim for a target of at least 80%, so there is something to be said for collecting data over a longer interval. What about a stepped wedge design? These are often presented as being statistically efficient. Figure 1B illustrates the classic stepped wedge scheme with a 'step-length', or interval between successive roll-outs, of 1 month. The power of this design is 91%—much more, in fact, than we need.

Now, a perceived advantage of the stepped wedge design is that all the sites end up receiving the intervention. But sites still have to wait: for the design in figure 1B the average wait is 5.5 months and the longest wait is 10 months. If this is unacceptable to sites then the design will fail. There are other designs with the same waiting characteristics: for the design

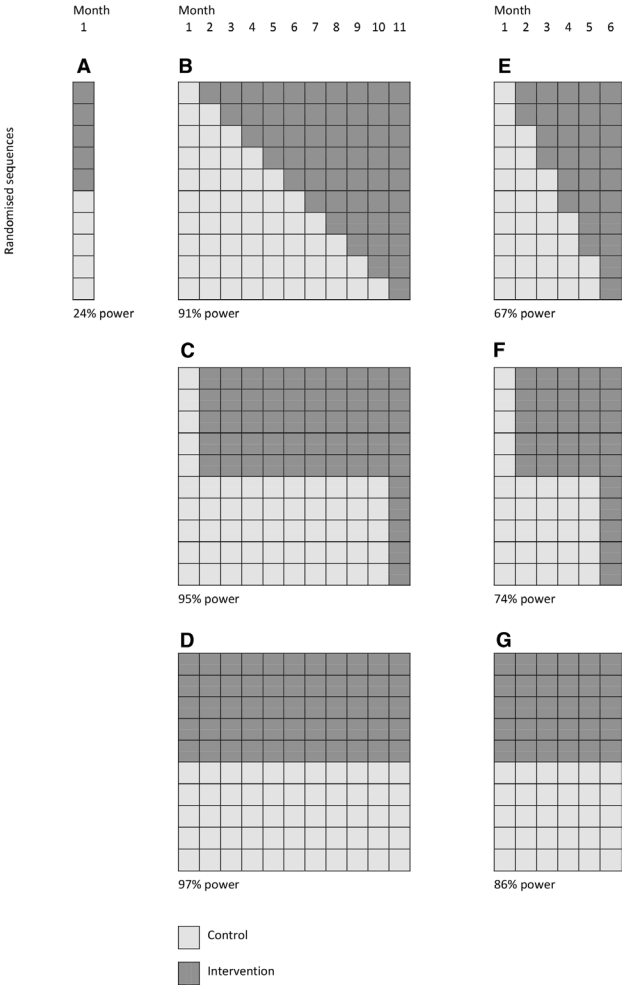

Figure 1 Designs for cluster randomised trials allowing crossover (in one direction) from a control to an intervention condition, either during or after the end of the trial, showing the statistical power of each design in a particular scenario (see box 1). Each row is a cluster and each column is a calendar month. Clusters are randomised to intervention sequences at the beginning of month 1. Design (A) runs for one month, Designs (B) to (D) run for 11 months, and Designs (E) to (G) run for 6 months. Designs (A), (D) and (G) are parallel group designs. Designs (B) and (E) are classic stepped wedge designs. Designs (C) and (F) randomise clusters to just two sequences, but have the same minimum, maximum and average waiting time for the intervention as the classic stepped wedge designs (B) and (E),

in figure 1C the average wait is again 5.5 months and the longest wait is 10 months. The latter design is simpler but does assume that several clusters can have the intervention implemented simultaneously. What may come as a surprise to some is that this simpler design has more power (95%) than the classic stepped wedge in the particular situation we are modelling—a phenomenon that arises, broadly speaking, when either the number of participants per cluster or the intracluster correlation (see box 1) is relatively small.[21 22]

If we go further, and abandon the idea that all clusters must begin in the control condition and end in the intervention condition, we arrive at the design in figure 1D, in which all the clusters are randomised to one condition or the other for the duration of the trial—that is, a 'parallel groups' design conducted over the same timescale as our stepped wedge design. This turns out to be the most statistically powerful design we have yet considered. Not all of the clusters receive the intervention within 10 months, but we do not have to leave things like that: we could have an agreement with sites to roll out the intervention to *all* of them immediately after the 11-month trial period, while we get on with analysing and publishing our results.

But what about that excess power? Could we get away with collecting *less* data? Figure 1E–G shows designs run over a 6-month interval, still divided into 1 month periods. This shows that we can achieve 86% power with a design that randomises half the clusters to the intervention for 6 months, and half to control (figure 1G). With a bit more tweaking it may be possible to uncover even more powerful alternative designs,[21 22] but this is not the point of the present exercise. The point is this: given 10 clusters and a step length of 1 month we might have jumped to the naïve conclusion that we should run a stepped wedge trial lasting 11 months. But this fixed idea would have prevented us from seeing in this instance that we could get the evidence we needed in a much shorter time and with a simpler design—randomising half the clusters to the intervention for 6 months, and half to control—with all sites then being free to receive the intervention (preferentially perhaps) or to go and seek it for themselves.

HOW WILL THE TRIAL BE ANALYSED?

So far, we have deliberately focused more on the design and conduct of stepped wedge trials than their analysis, but the two are connected and the latter generates just as much discussion. Combining quantitative information from between-site and within-site comparisons is relatively easy, although the methods that are commonly used—mixed regression and generalised estimating equations—rely heavily on statistical modelling.[23 24] Whether it is right to pursue complex modelling or to focus on more robust approaches to analysis is something methodologists continue to explore.[25–27] The challenges of data analysis should certainly not be ignored at the study design stage: simpler designs will present simpler analytical challenges.

One of the most important things when analysing a stepped wedge trial is to allow for the possibility of secular changes in outcomes over time (this is because time is confounded with treatment in a stepped wedge design). Yet we know from the work of others that this and other aspects of the analysis of stepped wedge trials are often handled inadequately in practice.[5 6] Concepts that seemed well defined, such as 'intention-to-treat' analysis,[28] become murkier: if the whole schedule for a stepped wedge trial slips by a month, do we still analyse according to the schedule we originally intended? Persuading clusters to comply with the precise schedule for crossover requires, in any case, a kind of 'extreme coordination'.[10 12] Stepped wedge designs also introduce new risks of bias.[29 30] In particular, the extended timescale may mean that individual participants are joining the study when the treatment condition is already known, leading to potential selection biases.

DISCUSSION

Stepped wedge designs provide a formal framework for evaluating interventions implemented at multiple sites. In this article we have focused on randomised evaluations, although non-randomised studies of interventions implemented at different times in different sites will share many of the features of stepped wedge trials.[31 32] The staggered implementation in a stepped wedge trial is also reminiscent of a series of Plan-Do-Study-Act (PDSA) cycles,[33 34] but the key difference is that the intervention remains the same in a stepped wedge trial. (Many stepped wedge trials might, incidentally, benefit from initial PDSA cycles to improve the intervention before the trial begins.)

Staggering the introduction of the intervention at different sites can offer statistical efficiency as well as practical benefits. But while efficiency and practicality may drive the choice of a stepped wedge design,[35] they can equally push you to consider alternatives. We recommend asking questions about the context for your research and seeking expert advice on design if needed, as it has not been possible for us to explore every design possibility in this article. Stepped wedge trials will undoubtedly continue to find widespread application, but they should not be seen as the solution to every evaluation problem in health services research or quality improvement, and in particular they are not the only way to ensure that everyone gets an intervention within a certain time frame. You should only extend the timescale of your evaluation and add complexity to the design (and consequently the analysis) because you have to, remembering that

there are also virtues in getting answers quickly and keeping things simple. Whether the stepped wedge is a cutting-edge tool or a blunt instrument depends entirely on how you use it.

Funding RH is a Senior Fellow with The Healthcare Improvement Studies (THIS) Institute. This Fellowship is funded by a grant from the Health Foundation to the University of Cambridge.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;6:54.
2 Mdege ND, Man M-S, Taylor Nee Brown CA, *et al*. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;64:936–48.
3 Beard E, Lewis JJ, Copas A, *et al*. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015;16:353.
4 Barker D, McElduff P, D'Este C, *et al*. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Med Res Methodol* 2016;16:69.
5 Martin J, Taljaard M, Girling A, *et al*. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open* 2016;6:e010166.
6 Grayling MJ, Wason JMS, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials* 2017;18:33.
7 Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28:182–91.
8 Hemming K, Taljaard M, McKenzie JE, *et al*. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ* 2018;363:k1614.
9 Kullgren JT, Krupka E, Schachter A, *et al*. Precommitting to choose wisely about low-value services: a stepped wedge cluster randomised trial. *BMJ Qual Saf* 2018;27:355–64.
10 Lenguerrand E, Winter C, Siassakos D, *et al*. Effect of hands-on interprofessional simulation training for local emergencies in Scotland: the THISTLE stepped-wedge design randomised controlled trial. *BMJ Qual Saf* 2020;29:122–34.
11 Hargreaves JR, Copas AJ, Beard E, *et al*. Five questions to consider before conducting a stepped wedge trial. *Trials* 2015;16:350.
12 Prost A, Binik A, Abubakar I, *et al*. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. *Trials* 2015;16:35.
13 International Committee of Medical Journal Editors. What is the ICMJE definition of a clinical trial? Available: http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/ [Accessed 24 Apr 2020].
14 Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
15 Hemming K, Haines TP, Chilton PJ, *et al*. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;350:h391.
16 Copas AJ, Lewis JJ, Thompson JA, *et al*. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015;16:352.
17 Snooks H, Bailey-Jones K, Burge-Jones D, *et al*. Effects and costs of implementing predictive risk stratification in primary care: a randomised stepped wedge trial. *BMJ Qual Saf* 2019;28:697–705.
18 Hooper R, Copas A. Stepped wedge trials with continuous recruitment require new ways of thinking. *J Clin Epidemiol* 2019;116:161–6.
19 Arnup SJ, McKenzie JE, Hemming K, *et al*. Understanding the cluster randomised crossover design: a graphical illustraton of the components of variation and a sample size tutorial. *Trials* 2017;18:381.
20 Spence J, Belley-Côté E, Lee SF, *et al*. The role of randomized cluster crossover trials for comparative effectiveness testing in anesthesia: design of the Benzodiazepine-Free cardiac anesthesia for reduction in postoperative delirium (B-free) trial. *Can J Anaesth* 2018;65:813–21.
21 Lawrie J, Carlin JB, Forbes AB. Optimal stepped wedge designs. *Stat Probab Lett* 2015;99:210–4.
22 Girling AJ, Hemming K. Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Stat Med* 2016;35:2149–66.
23 Hanley JA, Negassa A, Edwardes MDdeB, MDdeB E, *et al*. Statistical analysis of correlated data using generalized estimating equations: an orientation. *Am J Epidemiol* 2003;157:364–75.
24 Abel G, Elliott MN. Identifying and quantifying variation between healthcare organisations and geographical regions: using mixed-effects models. *BMJ Qual Saf* 2019;28:1032–8.
25 Thompson JA, Davey C, Fielding K, *et al*. Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Stat Med* 2018;37:2487–500.
26 Kasza J, Hemming K, Hooper R, *et al*. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res* 2019;28:703–16.
27 Kennedy-Shaffer L, de Gruttola V, Lipsitch M. Novel methods for the analysis of stepped wedge cluster randomized trials. *Stat Med* 2020;39:815–44.
28 Moher D, Hopewell S, Schulz KF, *et al*. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
29 Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ* 2009;339:b4006.
30 Zhan Z, van den Heuvel ER, Doornbos PM, *et al*. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. *J Clin Epidemiol* 2014;67:454–61.
31 Franklin BD, Reynolds M, Sadler S, *et al*. The effect of the electronic transmission of prescriptions on dispensing errors and prescription enhancements made in English community

pharmacies: a naturalistic stepped wedge study. *BMJ Qual Saf* 2014;23:629–38.
32 Bion J, Richardson A, Hibbert P, *et al*. 'Matching Michigan': a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England. *BMJ Qual Saf* 2013;22:110–23.
33 Reed JE, Card AJ. The problem with Plan-Do-Study-Act cycles. *BMJ Qual Saf* 2016;25:147–52.
34 Burke RE, Shojania KG. Rigorous evaluations of evolving interventions: can we have our cake and eat it too? *BMJ Qual Saf* 2018;27:254–7.
35 Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *Int J Epidemiol* 2020;94.
36 Kerry SM, Bland JM. The intracluster correlation coefficient in cluster randomisation. *BMJ* 1998;316:1455–60.
37 Cook JA, Julious SA, Sones W, *et al*. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ* 2018;363:k3750.
38 Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ* 2015;350:h2925.
39 Hooper R, Teerenstra S, de Hoop E, *et al*. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med* 2016;35:4718–28.
40 Hemming K, Kasza J, Hooper R, *et al*. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the shiny CRT calculator. *Int J Epidemiol* 2020;94. doi:10.1093/ije/dyz237. [Epub ahead of print: 22 Feb 2020].