Stepped-wedge cluster-randomised trials:

Haines, Terry P.; Hemming, Karla

DOI:
10.1016/j.jphys.2017.11.008

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Haines, TP & Hemming, K 2018, 'Stepped-wedge cluster-randomised trials: level of evidence, feasibility and reporting', *Journal of Physiotherapy*, vol. 64, no. 1, pp. 63-66. https://doi.org/10.1016/j.jphys.2017.11.008

Journal of Physiotherapy 64 (2018) 63–66

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Appraisal Research Note

Stepped-wedge cluster-randomised trials: level of evidence, feasibility and reporting

The stepped-wedge cluster-randomised trial is a form of cluster-crossover trial with unidirectional crossover between control and intervention conditions. It has become a popular research design, particularly in effectiveness and implementation of science research over the past decade.[1] The ascension of the stepped-wedge as a design of preference has been criticised in some quarters, yet praised in others.[2,3] However, there is a need for researchers and research users to better understand the intricacies of this design so that they can appropriately design, appraise and use information generated from trials using this approach. This Research Note describes the stepped-wedge design, discusses whether this design provides evidence on par with other randomised, controlled trial designs, and highlights some key considerations for feasibility and reporting.

The term stepped-wedge does not refer to a singular design, but to a family of designs. The stepped-wedge cluster design is usually characterised by four key features:

1) Clusters are exposed to both intervention and control conditions.
2) There is unidirectional crossover between control and intervention conditions.
3) Clusters transition between control and intervention conditions at different time points, the order of which is determined using a random process.
4) Outcome data are collected from each cluster for each time period in the study.

Beyond these features, there is a wide array of possible variations. Diagrammatic representations of some variations on the stepped-wedge design in comparison to related cluster-randomised designs have been previously provided.[3] Figure 1 expands on this by illustrating provision of control (white cells) and intervention (grey cells) conditions across time periods. The figure does not constitute an exhaustive catalogue of all possible variations on the stepped-wedge design. The figure shows a progression from a purely cross-sectional design (where different participants are measured at each time period, Figure 1 m) to a purely cohort design (where the same participants are measured repeatedly at each time period, Figure 1 p). Between these are two designs of note: 1) where individual participants are potentially present for more than one time period yet only provide one measurement (Figure 1 n); and 2) where individual subjects are potentially present for more than one time period yet provide a data point for each time period (Figure 1 o). Also within this figure are related versions of parallel cluster-randomised designs, the importance of which will be commented upon shortly.

Are stepped-wedge trials more at risk of bias than parallel-cluster randomised trials?

There is ongoing debate over the rigor of the stepped-wedge design and trustworthiness of this approach. Different positions have been adopted as to the calibre of evidence provided by stepped-wedge designs. Some authors have described stepped-wedge designs as 'quasi-experimental'[4] while other authors have described stepped-wedge designs as 'experimental',[5] although a clear justification was not articulated in either case. Whether a stepped-wedge trial might be at greater risk of bias than its closest experimental comparator – the parallel, cluster-randomised, controlled trial – will now be discussed (Figure 1 a to d).

A key difference between the stepped-wedge and the parallel cluster-randomised trial that creates potential for the stepped-wedge to produce more biased results is that there is an unequal distribution of control and intervention periods over calendar time. This is particularly important if there are secular trends associated with the study outcome. For example, medical hospital admissions in New Zealand between April 1993 and September 2008 were nearly 20% lower in February (summer) than August (winter).[6] A 7-month stepped-wedge study seeking to test an intervention that reduces this outcome and begins with a period of all clusters in a control period in August and ending with all clusters being in an intervention period in February will naturally bias the outcome in favour of the intervention. A similar problem arises for outcomes that naturally improve or diminish over time (maturation effects).

Stepped-wedge studies therefore must adjust for time effects to ensure that the estimated treatment effect is not confounded. It is not recommended to statistically test for the occurrence of time effects and then decide whether to adjust or not.[7] The accuracy of the resulting estimated treatment effect will be dependent on the model's ability to capture these time effects. Treatment effects estimated from a model in which the time effects have been mis-specified are likely to be biased or have confidence intervals that are too wide or narrow.[8] Alternatives, such as designing the study to be free of known secular patterns (such as creating a balance on seasonal effects), whilst likely to reduce the impact of any time effects cannot rule out other confounding effects resulting from the natural imbalance with respect to time the study creates.

Whilst the stepped-wedge design induces a risk of bias due to time effects (albeit one which in theory can be adjusted for), the design does induce some possible advantages. In a stepped-wedge study the advantage is that each cluster contributes information to both the intervention and control conditions, whereas in a parallel cluster-randomised trial the clusters contribute data to only the control or intervention condition. Underlying differences between clusters can confound the results arising from a parallel cluster-randomised trial, particularly if the number of clusters is small and variability between clusters is high. This is likely an important consideration in many fields. For example, the field of falls prevention among hospital inpatients often uses interventions delivered across an entire ward, with randomisation of wards inherent to the study design. Different ward types may have better pre-existing approaches to falls prevention, and some have been found to have a different propensity to report falls on incident reports.[9] Thus, the results from a parallel cluster-randomised study

https://doi.org/10.1016/j.jphys.2017.11.008

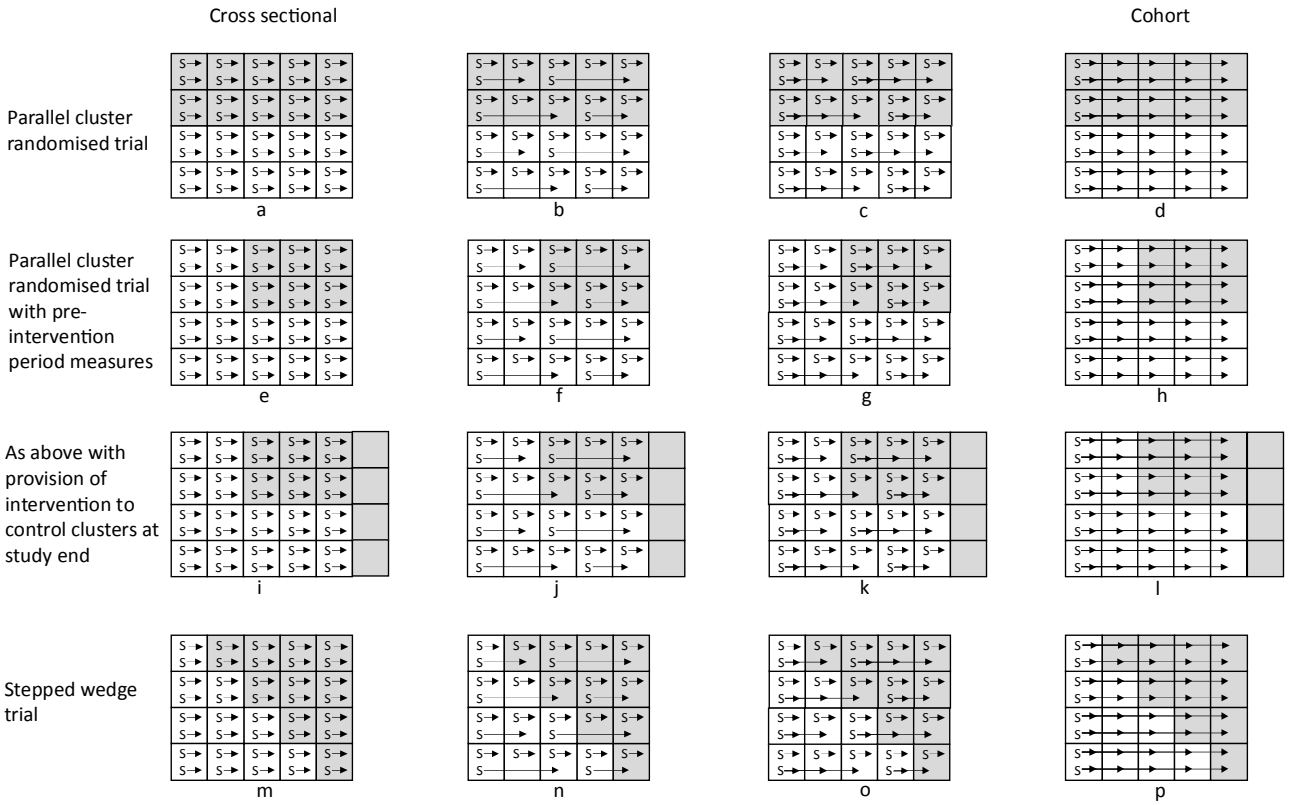

Figure 1. Variants of stepped-wedge and parallel, cluster-randomised designs.

that randomises a small number of wards to intervention and control groups could be easily confounded by these differences through 'unlucky' randomisation. Stepped-wedge trials are likely to be less affected by these underlying differences between clusters, as each cluster contributes to both intervention and control conditions.[3]

Proponents of parallel cluster-randomised trials in preference to stepped-wedge could argue at this point that modifications to the parallel cluster-randomised trial could be employed. Some form of restricted randomisation procedure could be used to increase the likelihood of a balance on measured cluster level covariates across intervention and control condition arms. These methods include matching, stratification, covariate constrained randomisation and minimisation (when the clusters are not all enrolled at the time of randomisation).[10] However, none of these methods ensure a balanced design on measured factors, and none can ensure a balanced design on unmeasured confounders.[11] Potentially confounding variables can be statistically adjusted for, although these procedures do not account for imbalance on any characteristics that are not measured.[12]

A parallel cluster-randomised trial with a pre-intervention period measure could also be used. This is depicted in Figure 1 e to h. The advantage of this approach is that inter-cluster variability in the outcome measure that can lead to baseline imbalances can be explicitly modelled and accounted for in the analyses. There are a range of ways this can be done. Typical analysis methods include adjustment for a cluster-level mean or individual-level value of outcome at baseline.[13] However, an alternative is adjustment for fixed effect for time period of measurement, which is on a par with how the stepped-wedge design is analysed.[14] Whilst using pre-intervention period measures can provide advantages in terms of increased statistical power, the design in common with the stepped-wedge design requires model-based methods to adjust for confounders (either adjustment for the baseline value of the outcome or the time effect, dependent on model choice) and so has the potential to provide either biased estimates of treatment effects or biased standard errors in the case of mis-specification of model forms.[8]

A parallel cluster-randomised trial with a pre-intervention period measure was used to understand the impact of providing low hospital beds for the prevention of falls across 18 hospital wards over a 12-month period.[13] An interesting element of this study was that the low beds were provided to the control wards following study conclusion. This strategy, sometimes referred to as a waiting lists design, can be used by investigators to help promote recruitment of clusters. If outcomes had been collected following provision of the intervention to the control wards and used in the analysis, this parallel cluster-randomised controlled experiment would have evolved into a stepped-wedge trial. This evolution of designs depicted in Figure 1 could be considered analogous to a 'March of Progress'[15] from parallel cluster-randomised trials through to stepped-wedge in terms of ability to mitigate cluster-level baseline imbalances. If readers are prepared to accept the parallel cluster-randomised trial with pre-intervention period measures as a means for handling potential baseline imbalance, then it could be argued that they should also be willing to accept the stepped-wedge as an extension of this design.

Both stepped-wedge and cluster-randomised trials with a pre-intervention period measure create the potential for within-participant contamination of control and intervention condition exposure. This can arise when measurements are taken for participants who are recruited under the control condition but continue to be exposed after the cluster has transitioned to the intervention condition (Figure 1 f, j, n). This situation may arise in a hospital setting where a particular intervention is delivered sequentially across participating wards, and patient length of stay

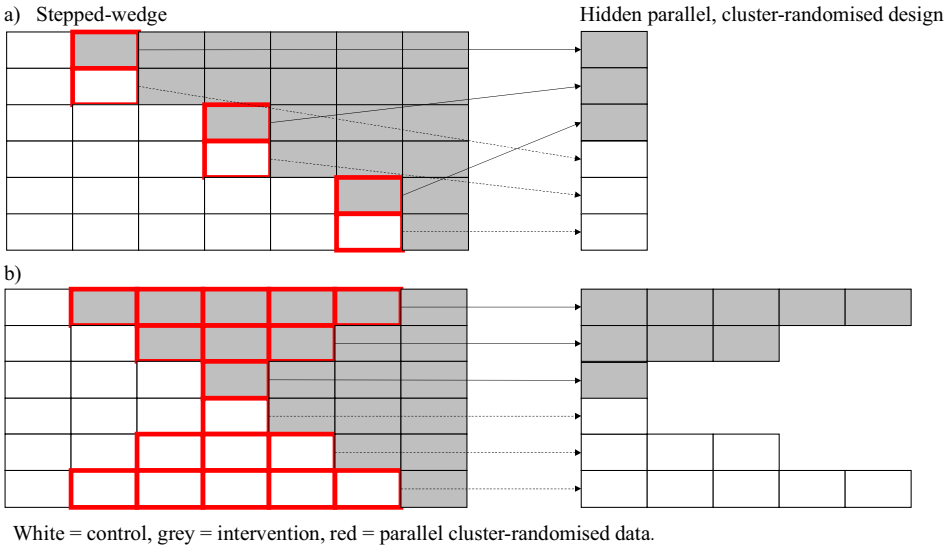

Figure 2. Data from parallel, cluster-randomised trials hidden within the stepped-wedge design by pairing subsets of a) adjacent clusters within the same time periods, and b) clusters from opposite sides of the design.

in hospital is the outcome of interest. Wash-out periods between control and intervention periods within stepped-wedge designs have previously been advocated for allowing time for interventions to be implemented and clinical effects to manifest,[3] but could also be applied in this situation to help deal with this issue.

A final observation on this issue is that data from a parallel cluster-randomised trial design lies within each stepped-wedge. This observation is illustrated (Figure 2). Here adjacent pairs of clusters in the stepped-wedge design contribute control and intervention group data at different time points across the study. This is what would have occurred had a parallel cluster trial design been implemented with different starting time points for pairs of clusters. A disadvantage of the resultant dataset that can be extracted from the stepped-wedge design is the loss of data and loss of power that would result. The advantage is that it is not necessary to statistically adjust for time periods to account for secular trends or maturation effects. Figure 2 also demonstrates how the same principle can be applied to pairs of clusters at the opposite sides of the stepped-wedge design, so that a maximum amount of data can be extracted and still be consistent with the parallel, cluster-randomised trial design. The resultant dataset, which is described as the 'within-wedge' analysis, has a characteristic shape of balancing pyramids (one regular, one inverted on top).

Feasibility of stepped-wedge trials

Previous criticisms of stepped-wedge trials have tended not to focus on risks of bias, as discussed above. Rather they have largely focused on issues of feasibility, cost and burden to participants.[2] Much of this has been drawn from the need to collect outcomes in every time period of the study. What might have required only one measurement per participant in a parallel cluster-randomised trial becomes several in a stepped-wedge trial if it uses a cohort design (Figure 1 d, h, l, p). The burden on an individual participant remains of particular concern for the ethical conduct of research. Contexts amenable to cross-sectional versions of the designs will have less burden on individual participants; contexts where study outcomes are measured as a part of routine care may have fewer issues regarding feasibility and participant burden when using the stepped-wedge design.

A feature that affects the feasibility of stepped-wedge trials is the necessity for interventions to be delivered at specific time points. It is conceivable that there are many interventions that would be subject to uncertain timing and delays that are outside the control of researchers.[16] Furthermore, study sites would need to be ready to commence data collection at the same time in a stepped-wedge. A parallel cluster-randomised trial design with staggered starting times that groups clusters to commence data collection might be a more feasible design for these situations.[17]

Stepped-wedge designs might increase study duration compared to parallel cluster-randomised designs, and so increase the risk that a range of threats to the trial may interfere with its conduct. For example, two stepped-wedge trials were undertaken back-to-back across two hospitals in Melbourne, Australia, over a combined 16-month calendar period.[18] However, after study commencement, one of these hospitals underwent a restructure that led to the closure of one of the study wards. Parallel cluster-randomised trials may potentially enable a research team to get in and out of a research location faster and reduce the opportunity for unanticipated problems to interfere with the trial.

Reporting in stepped-wedge trials

Several aspects require special consideration when reporting a stepped-wedge cluster-randomised trial. First, the CONSORT Guidelines for reporting of cluster-randomised trials recommends that summary values for both treatment and control conditions be reported in abstracts.[19] However, in a stepped-wedge trial this summative data may show a distorted picture of the effect of the intervention due to the confounding effect of time. For readers of abstracts, it is important to note that the summative data in stepped-wedge designs may not correspond particularly well with the treatment effect size estimates generated, although it is the latter that should be paid heed to.

Second, interpretation of findings from stepped-wedge designs using line graphs could be aided greatly through the use of line graphs tracing cluster outcomes over time. Line graphs where outcomes from clusters all commence at the same calendar time may be particularly useful for identifying secular trends in the data. This method is illustrated in Figure 3 a, and shows outcomes (proportion of patients who experience an adverse event) across six clusters. Figure 3 shows a possible secular trend with an increase in patients experiencing adverse events in February being of concern (Figure 3 a).[18] This visual presentation of data is less useful for understanding the effect of the intervention, as readers cannot tell when each cluster commenced the intervention. Previous authors have attempted to address this issue by marking transition points based on calendar time using annotations or symbols on individual lines.[20] An alternative approach may be to

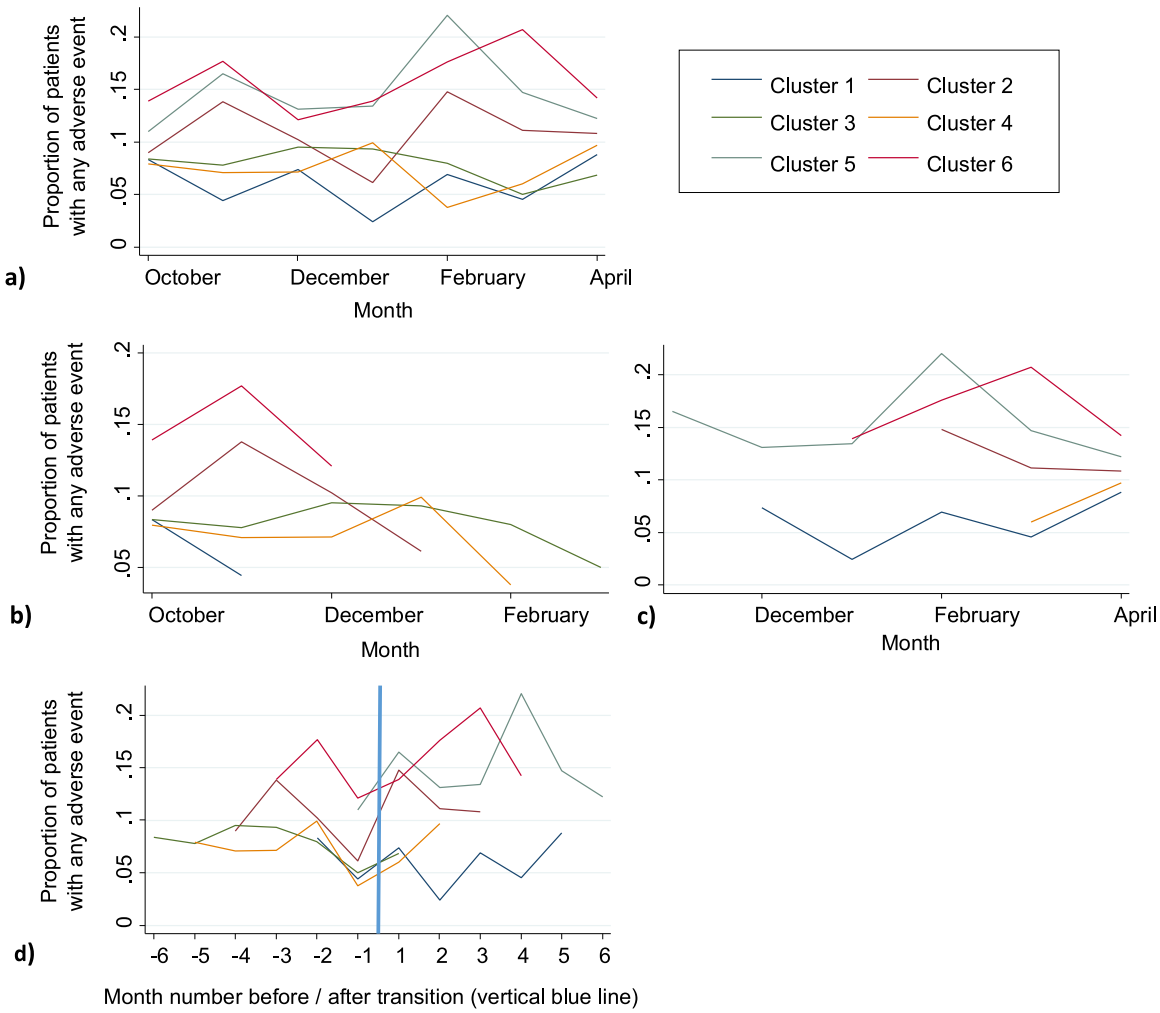

Figure 3. Line graphs based upon: a) whole trial data using calendar time; b) control period data using calendar time; c) intervention period data using calendar time; and d) whole trial data using time relative to the transition period.

present two graphs aligned with calendar time but separated into data from control and intervention periods (Figures 3 b and c). Another option would be to produce a line graph where cluster outcomes are aligned at the time when clusters transition from control to intervention periods (Figure 3 d). The 'transition-relative' graph can easily identify data points attributable to the intervention and control periods, which are difficult to distinguish using the conventional approach. However, this approach masks secular trends and could be visually misleading if they are present. It may be that a combination of graphs is most illustrative.

Conclusions

Stepped-wedge trials belong to a broad family of experimental research designs that can provide opportunities to undertake rigorous research at low cost in contexts where outcome data are routinely recorded and easily accessed. These designs have both methodological advantages and disadvantages compared to other cluster randomised designs, but should arguably hold a comparable position in evidence hierarchies. There are some concerns about the feasibility of stepped-wedge designs, particularly where the research team does not directly control elements of the intervention. Researchers should ensure that accurately reported results reflect the effect of the intervention after adjustment for the confounding effect of time.

Acknowledgements: Nil.

Competing interests: Nil.

Provenance: Invited. Peer reviewed.

Correspondence: Terry Haines, School of Primary and Allied Health Care, Monash University, Australia. Email: Terrence.haines@monash.edu

Terry P Haines[a,b] and Karla Hemming[b]

[a]*School of Primary and Allied Health Care, Monash University, Frankston, Australia*

[b]*Institute of Allied Health Research, University of Birmingham, Birmingham, UK*

References

1. Beard E, et al. *Trials.* 2015;16:353.
2. Kotz D, et al. *J Clin Epidemiol.* 2012;65:1249–1252.
3. Hemming K, et al. *BMJ.* 2015;350:h391.
4. Handley MA, et al. *J Am Board Fam Med.* 2011;24:589–596.
5. Torgerson CJ, et al. *Handbook of Practical Program Evaluation*. 144.
6. Jones R. *Med J Aust.* 2010;193:298–300.
7. Hemming K, et al. *Trials.* 2017;18:101.
8. Thompson JA, et al. *Stat Med.* 2017;36(23):3670–3682.
9. Haines T, et al. *J Am Geriatr Soc.* 2009;57:517–523.
10. Ivers NM, et al. *Trials.* 2012;13:120.
11. Hayes RJ, Moulton LH. *Cluster randomised trials.* 2nd Ed. Florida: Chapman & Hall; 2017.
12. Assman S, et al. *Lancet.* 2000;355(9209):1064–1069.
13. Klar N, Darlington G. *Stat Med.* 2004;23(15):2341–2357.
14. Coffman CJ, et al. *BMJ Open.* 2016;6:e013096.
15. Shelley C. *Informal Logic.* 2001;21(2).
16. Prost A, et al. *Trials.* 2015;16:351.
17. Hemming K, et al. *Stat Med.* 2015;34:181–196.
18. Haines TP, et al. *PLoS Med.* 2017;14:e1002412.
19. Campbell MK, et al. *BMJ.* 2012;345:e5661.
20. Liddy C, et al. *Implement Sci.* 2015;10:150.