Harvard University
Harvard University Biostatistics Working Paper Series

Year 2018 *Paper* 216

Power Calculation for Cross-Sectional Stepped Wedge Cluster-Randomized Trials with Variable Cluster Sizes

Linda J. Harrison* Tom Chen†
Rui Wang‡

*Harvard University, ljh916@g.harvard.edu
†Harvard Pilgrim Health Care Institute, tomchen00@gmail.com
‡Harvard University, rwang@hsph.harvard.edu

Power Calculation for Cross-Sectional Stepped Wedge Cluster-Randomized Trials with Variable Cluster Sizes

Linda J. Harrison, Tom Chen, and Rui Wang

Abstract

Standard sample size calculation formulas for Stepped Wedge Cluster Randomized Trials (SW-CRTs) assume that cluster sizes are equal. When cluster sizes vary substantially, ignoring this variation may lead to an under-powered study. We investigate the relative efficiency of a SW-CRT with varying cluster sizes to equal cluster sizes, and derive variance estimators for the intervention effect that account for this variation under the assumption of a mixed effects model; a commonly-used approach for analyzing data from cluster randomized trials. When cluster sizes vary, the power of a SW-CRT depends on the order in which clusters receive the intervention, which is determined through randomization. We first derive a variance formula that corresponds to any particular realization of the randomized sequence and propose efficient algorithms to identify upper and lower bounds of the power. We then obtain an "expected" power based on a first-order approximation to the variance formula, where the expectation is taken with respect to all possible randomization sequences. Finally, we provide a variance formula for more general settings where only the mean and coefficient of variation of cluster sizes, instead of exact cluster sizes, are known in the design stage. We evaluate our methods through simulations and illustrate that the power of a SW-CRT decreases as the variation in cluster sizes increases, and the impact is largest when the number of clusters is small.

Biometrics 64, 1–23 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2018

Power Calculation for Cross-Sectional Stepped Wedge Cluster Randomized Trials with Variable Cluster Sizes

Linda J Harrison[1,*], Tom Chen[1,2], and Rui Wang[1,2]

[1]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, U.S.A

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, U.S.A.

email: ljh916@g.harvard.edu

Summary: Standard sample size calculation formulas for Stepped Wedge Cluster Randomized Trials (SW-CRTs) assume that cluster sizes are equal. When cluster sizes vary substantially, ignoring this variation may lead to an underpowered study. We investigate the relative efficiency of a SW-CRT with varying cluster sizes to equal cluster sizes, and derive variance estimators for the intervention effect that account for this variation under the assumption of a mixed effects model; a commonly-used approach for analyzing data from cluster randomized trials. When cluster sizes vary, the power of a SW-CRT depends on the order in which clusters receive the intervention, which is determined through randomization. We first derive a variance formula that corresponds to any particular realization of the randomized sequence and propose efficient algorithms to identify upper and lower bounds of the power. We then obtain an "expected" power based on a first-order approximation to the variance formula, where the expectation is taken with respect to all possible randomization sequences. Finally, we provide a variance formula for more general settings where only the mean and coefficient of variation of cluster sizes, instead of exact cluster sizes, are known in the design stage. We evaluate our methods through simulations and illustrate that the power of a SW-CRT decreases as the variation in cluster sizes increases, and the impact is largest when the number of clusters is small.

Key words: Stepped Wedge (SW); Cluster Randomized Trial (CRT); power; sample size; cluster size variation; cross-sectional

1. Introduction

Cluster Randomized Trials (CRTs) are studies in which clusters of individuals, rather than individuals themselves, are randomized to intervention groups (Donner and Klar, 2000; Hayes and Moulton, 2009). In a Stepped Wedge Cluster Randomized Trial (SW-CRT), clusters are randomized to cross forward from control to intervention at certain time points commonly termed as steps (Brown and Lilford, 2006; Hemming et al., 2015). The response is measured between each step; either on the same individuals each time (cohort SW-CRT) or on different individuals (cross-sectional SW-CRT). Table 1 provides a schematic for an example SW-CRT. The popularity of SW-CRTs has increased in recent years and the design is often implemented when a stepwise intervention initiation offers a practical solution.

The sample size required for a standard CRT is inflated compared with an individually randomized trial because outcomes of participants from the same cluster are correlated. A commonly-used inflation factor, also known as the design effect (DE), is $[1 + (n - 1)\rho]$ where n is the mean cluster size and ρ the intra-cluster correlation. Many articles (Kerry and Bland, 2001; Manatunga et al., 2001; Hoover, 2002; Eldridge et al., 2006; van Breukelen et al., 2007) have investigated the impact of cluster size variation on the power and sample size of standard CRTs, and demonstrated that cluster size variation results in lower power. A simple adjustment to the DE for standard CRTs (Eldridge et al., 2006) to account for variability in cluster sizes is given by $[1 + (n(1 + CV^2) - 1)\rho]$ where CV is the coefficient of variation; the ratio of the standard deviation of cluster sizes to the mean cluster size. As an example, a cluster size CV of 0.7 with a mean cluster size of $n = 100$ and an intra-cluster correlation of $\rho = 0.05$, results in sample size increases of 41%.

A commonly-used approach to analyze data from a cross-sectional SW-CRT is based on a linear mixed effects model (Hussey and Hughes, 2007). A DE based on such a model that depends on equal cluster size (n), the intra-cluster correlation (ρ), the number of steps,

and the number of time-points each cluster contributes n samples at baseline and between each step has been proposed (Woertman et al., 2013). Although equal cluster sizes (n) are desirable to achieve statistical efficiency, it may be infeasible logistically or non-optimal from the sampling perspective. In some settings, the design may be to sample a fixed percentage of the population, which naturally leads to unequal cluster sizes due to varying community sizes and differential rate of participant refusal. A recent systematic review of 101 SW-CRTs detected that 48% of studies included clusters that were known to vary in size (Kristunas et al., 2017). Some cross-sectional SW-CRTs have clusters that are inherently variable in size. For example, in a study of fall rates in hospital rehabilitation units (Poldervaart et al., 2013, 2017), the number of patients enrolled per unit was limited by the number of beds per ward, ranging from 14 to 90. Similarly, in the HEART impact trial (Hill et al., 2014, 2015) the goal was to recruit all eligible patients with chest pain, so the number recruited per department varied with hospital size. While it may be possible to limit the recruitment to the first M consenting participants from each site as in Bashour et al. (2013), this would result in a prolonged recruitment period for those sites with a smaller candidate pool. It would be useful to assess whether or not allowing cluster sizes to vary by recruiting more participants in larger clusters would lead to a more feasible study design with similar power even though the overall sample size required may be larger.

The impact of ignoring variation in cluster sizes on sample size calculation in SW-CRTs is unclear. Through simulation studies, Kristunas et al. (2017) reported that a variation in cluster size did not lead to notable loss of power in cross-sectional SW-CRTs with continuous outcomes, but the authors noted they had only examined a small range of parameters. Martin et al. (2018) calculated the power of a cross-sectional SW-CRT with unequal cluster sizes by numerically inverting the precision matrix, and further noted that the power depended on the order in which the variable size clusters were randomized to initiate the intervention.

Martin et al. (2018) went on to utilize simulation methods with a gamma distribution for cluster sizes to estimate the median and quartiles of the power across different randomization sequences. Matthews (2016) targets the question of whether an optimal or near-optimal ordering in which the variable size clusters initiate the intervention to achieve high power can be determined without an exhaustive search across all possible randomization sequences. He proposes a solution for the special cases where the intra-cluster correlation is extremely large or small. Girling (2018) recently derived an analytical formula for the relative efficiency (RE) of a SW-CRT with unequal compared to equal cluster sizes under a constrained randomization setting. The derivation relies on several clusters being randomized at each step, such that by stratifying the randomization procedure there is no inequality in the total size of all the clusters randomized to initiate the intervention at each step. To the best of our knowledge, analytical formulas for sample size and power estimation for cross-sectional SW-CRTs in general settings with varying cluster sizes have not been derived.

Under a linear mixed effects model framework, in Section 2 we derive three analytical formulas of variance estimates for power calculations that account for cluster size variation in cross-sectional SW-CRTs. The first assumes cluster sizes and their order of randomization are both known, and allows us to identify upper and lower bounds for the power of a SW-CRT (Section 2.2 and 2.3). The second provides a closed form expression for the expected variance before randomization when all cluster sizes are known (Section 2.4), and the third approximates this value if only the mean and CV of cluster size can be estimated in the planning stages of a SW-CRT (Section 2.5). In Section 2.6, we derive the DE for cross-sectional SW-CRTs as compared to individually randomized trials accounting for varying cluster sizes, a correction factor (CF) to correct sample size calculation and the expected relative efficiency (RE) of a cross-sectional SW-CRT with unequal compared to equal cluster sizes. Our simulation study and its results are described in Section 3, and illustrative

examples on the impact of cluster size variation on the sample size and RE of a cross-sectional SW-CRT are provided in Section 4. We end the paper with a Discussion (Section 5).

[Table 1 about here.]

2. Methods

2.1 *Notation and Model*

Henceforth, we consider a cross-sectional design with individuals $k = 1, \cdots, n_i$ sampled from cluster $i = 1, \cdots, I$ at every time-point $j = 1, \cdots, T$. The landmark paper on SW-CRTs by Hussey and Hughes (2007) proposes the following model for response variable Y_{ijk}:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + X_{ij}\theta + e_{ijk}, \qquad \alpha_i \sim N(0, \tau^2),\ e_{ijk} \sim N(0, \sigma_e^2)$$

where α_i is the random effect for cluster i, β_j is a fixed effect corresponding to time j ($j = 1, ..., T-1$ with $\beta_T = 0$ for identifiability), X_{ij} is an indicator of treatment (1 = intervention, 0 = control) in cluster i at time j, θ is the treatment effect, and e_{ijk} is the subject-specific error independent of α_i.

2.2 *Treatment Effect Variance when Cluster Sizes and Order of Randomization Known*

Under the linear mixed effects model, estimates for the fixed effects can be obtained using weighted least squares (WLS). Let Z be the $IT \times (T+1)$ design matrix corresponding to the parameter vector $\eta = (\mu, \beta_1, \beta_2, ..., \beta_{T-1}, \theta)$. Then, the WLS estimator is $\hat{\eta} = (Z^T V^{-1} Z)^{-1}(Z^T V^{-1} Y)$ and the treatment effect denoted by $\hat{\theta}$ is the $(T+1)$th element of $\hat{\eta}$. The covariance matrix of $\hat{\eta}$ is $(Z^T V^{-1} Z)^{-1}$, where V is a block diagonal matrix provided in Web Appendix A.

To test the hypothesis $H_0 : \theta = 0$ versus $\theta = \theta_A$, we can use a Wald test based on $W = \hat{\theta} \Big/ \sqrt{\text{Var}(\hat{\theta})}$ where $\hat{\theta}$ is the estimated treatment effect from WLS. The approximate

power for conducting a two-tailed test of size α is:

$$1-\beta \approx \Phi\left(\frac{\theta_A}{\sqrt{\text{Var}(\hat{\theta})}} - z_{1-\alpha/2}\right)$$

where β is the probability of a type II error, $1-\beta$ is the statistical power, Φ is the cumulative standard normal distribution function, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution function, and $\text{Var}(\hat{\theta})$ is the $(T+1),(T+1)$ element of the covariance matrix $(Z^T V^{-1} Z)^{-1}$

A closed form expression for the treatment effect variance given one particular realization of the randomization sequence can be derived (See Web Appendix A for details):

$$\text{Var}(\hat{\theta}|P=p) = \frac{fT(f+gT)}{fT(f+gT)(\ell-z) - (f+gT)y^2 - f(Tw-\ell^2)} \tag{1}$$

where

$$f = \sum_{i=1}^{I} \frac{1}{\sigma_i^2 + T\tau^2}, \qquad g = \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)}, \qquad \ell = \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2},$$

$$z = \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \left(\sum_{j=1}^{T} X_{ij}\right)^2, \quad y = \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2 + T\tau^2}, \quad w = \sum_{j=1}^{T} \left(\sum_{i=1}^{I} \frac{X_{ij}}{\sigma_i^2}\right)^2$$

and $\sigma_i^2 = \sigma_e^2/n_i$. When the sample size is the same for all clusters this variance simplifies to the variance provided in equation 8 of Hussey and Hughes (2007) (See Web Appendix B).

Inspection of the terms in the variance formula given in equation 1 reveals that components of the denominator (ℓ, z, y^2, $Tw-\ell^2$) depend upon the order in which the clusters are randomized to intervention; $P=p$ in equation 1 denotes a particular realization of the randomization sequence. For example, $\ell = \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2} = \frac{1}{\sigma_e^2}(n_1(X_{11}+...+X_{1T})+...+n_I(X_{I1}+...+X_{IT}))$ will be larger if a large cluster is randomized to intervention first, i.e. if n_1 is large. This is because in SW-CRTs the first cluster randomized receives treatment for the most time periods, so as X_{ij} is the indicator of treatment status, $X_{11}+...+X_{1T}$ will be greater than or equal to the other summations. If randomization of all clusters in a SW-CRT has taken place, this variance formula could be utilized to check the power after randomization. If the power is too low, the formula could be used to determine the number of additional measurements

needed to increase the power. As it is customary to calculate the power in planning stages of a trial before randomization, in Section 2.3 and 2.4 we provide a method to find the upper and lower bound of the treatment effect variance across all possible randomizations, as well as a closed form expression for the expected value of the variance to obtain the study power in an average sense.

2.3 *Method to Find Upper and Lower Bounds for the Treatment Effect Variance*

To find the upper and lower bounds for the power across all possible randomizations, the maximum and minimum of the denominator $D(v_1, \cdots, v_I)$ of the variance of the treatment effect estimator in equation 1 was sought. This requires optimizing $D(v_1, \cdots, v_I)$ over the order of varying cluster sizes; that is,

$$\begin{aligned} &\text{minimize/maximize: } D(v_1, \cdots, v_I) \\ &\qquad \text{subject to: } (v_1, \cdots, v_I) \text{ is a permutation of } (n_1, \cdots, n_I) \end{aligned}$$

The above optimization problem is classified as an assignment problem with a quadratic objective, since $D(v_1, \cdots, v_I)$ is quadratic in $v_1, \cdots, v_I$. In general, an assignment problem with a non-linear objective is not only NP-hard (i.e. cannot be solved in polynomial time), but also does not have a "good" algorithm to solve it other than cycling through all $I!$ permutations. Fortunately, a reparametrization involving permutation matrices reformulates the problem into a mixed-integer quadratic programming (MIQP) problem, which, while still NP-hard, has excellent algorithms to solve it. Specifically, the reparametrization (see Web Appendix C) takes the form

$$\begin{aligned} &\text{minimize/maximize: } R^T M R + D^T R \\ &\text{subject to: } \sum_{i=1}^{I} R_{(s-1)I+i} = 1 \quad \forall s = 1, \cdots, I \\ &\qquad\qquad \sum_{s=1}^{I} R_{(s-1)I+i} = 1 \quad \forall i = 1, \cdots, I \end{aligned}$$

$$R_{(s-1)I+i} \in \{0, 1\}$$

where

- R is a vector of length I^2 decision variables
- M is an $I^2 \times I^2$ matrix with elements

$$M_{(s-1)I+i,(t-1)I+j} = -(f + gT)\left(\sum_{k=1}^{I} X_{sk}\right)\left(\sum_{k=1}^{I} X_{tk}\right)\frac{1}{(\sigma_i^2 + \tau^2 T)(\sigma_j^2 + \tau^2 T)} - f\left[T\sum_{k=1}^{I} X_{sk}X_{tk} - \left(\sum_{k=1}^{I} X_{sk}\right)\left(\sum_{k=1}^{I} X_{tk}\right)\right]\frac{1}{\sigma_i^2\sigma_j^2}$$

  for $s, t, i, j = 1, \cdots, I$

- D is a vector of length I^2 with elements

$$D_{(t-1)I+j} = fT(f + gT)\left[\left(\sum_{k=1}^{I} X_{tk}\right)\frac{1}{\sigma_j^2} - \left(\sum_{k=1}^{I} X_{tk}\right)^2 \frac{\tau^2}{\sigma_j^2(\sigma_j^2 + \tau^2 T)}\right]$$

  for $t, j = 1, \cdots I$

The above form can now be solved by algorithms implemented by solvers such as Gurobi (2018). The decision variable vector R is a vectorization of the permutation matrix which encodes the order clusters should be placed in the randomization sequence to obtain highest or lowest possible power. For example, suppose there are four clusters of size 10, 15, 45 and 50. An optimal solution $R = (0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1)$ is a vectorization of the matrix

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ with optimal order } \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 10 \\ 15 \\ 45 \\ 50 \end{bmatrix} = \begin{bmatrix} 45 \\ 15 \\ 10 \\ 50 \end{bmatrix}$$

We require the vectorized form R in order to feed the decision variables into a solver. One sequence that optimizes the objective function will be found using this method and the maximum/minimum power of the SW-CRT design can be obtained by using this optimal sequence to first calculate the treatment effect variance using equation 1 and then to estimate

the power of the Wald test using this variance. We note that the randomization sequence achieving highest or lowest power can be non-unique. In the example, a sequence of 50, 10, 15 and 45 would also attain the same maximum power. An exhaustive search over all the possible randomizations could be conducted to identify all sequences of the cluster sizes that have higher power. However, for design purposes, we are most interested in the best and worst case scenarios for the power, identifying all sequences that are associated with each extreme case is not necessary.

2.4 *Expected Treatment Effect Variance when Cluster Sizes Known*

When all the cluster sizes are known prior to randomization, we derive a closed form expression for first order approximation of the expected value of $\text{Var}(\hat{\theta}|P)$, where the expectation is taken across all possible randomization realizations.

To proceed, we consider the settings described in Woertman et al. (2013) where q clusters are randomized at each step, where q is any divisor of the total number of clusters I, where each cluster contributes samples at b baseline time-points before any cluster begins the intervention and at t time-points after each step. Hence, cross-sectional samples from the clusters will be taken at $T = (I/q)t + b$ time-points. Such balanced designs are commonly-used in practice. If $K = I/q$ represents the number of steps, then T the total time-points can also be written $Kt + b$. An an example, a design corresponding to $b = 1,\ t = 2,\ q = 2$ results in the following treatment status matrix, where X_{ij} is the indicator of treatment

(1=intervention, 0=control) in cluster i at time j:

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} & X_{14} & X_{15} & X_{16} & \cdots & X_{1(T-1)} & X_{1T} \\ X_{21} & X_{22} & X_{23} & X_{24} & X_{25} & X_{26} & \cdots & X_{2(T-1)} & X_{2T} \\ X_{31} & X_{22} & X_{33} & X_{34} & X_{35} & X_{36} & \cdots & X_{3(T-1)} & X_{3T} \\ X_{41} & X_{22} & X_{43} & X_{44} & X_{45} & X_{46} & \cdots & X_{4(T-1)} & X_{4T} \\ X_{51} & X_{22} & X_{53} & X_{54} & X_{55} & X_{56} & \cdots & X_{5(T-1)} & X_{5T} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{I1} & X_{I2} & X_{I3} & X_{I4} & X_{I5} & X_{I6} & \cdots & X_{I(T-1)} & X_{IT} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}$$

In the variance formula given in equation 1, three terms in the denominator, $(\ell - z)$, (y^2) and $(Tw - \ell^2)$ depend on the order clusters are randomly assigned to initiate the intervention. To calculate the three expectations, $\mathbb{E}(\ell - z)$, $\mathbb{E}(y^2)$ and $\mathbb{E}(Tw - \ell^2)$, we derived $\mathbb{P}(X_{ik} = 1)$, the probability cluster i is treated at time-point k, and $\mathbb{P}(X_{ik} = 1, X_{jl} = 1)$, the probability cluster i is treated at time-point k and cluster j is treated at time-point l, as follows (Web Appendix D):

$$\mathbb{P}(X_{ik} = 1) = \frac{\lceil \frac{k-b}{t} \rceil q}{I}$$

$$\mathbb{P}(X_{ik} = 1, X_{jl} = 1) = \frac{(\lceil \frac{k-b}{t} \rceil \wedge \lceil \frac{l-b}{t} \rceil)q}{I} \frac{(\lceil \frac{k-b}{t} \rceil \vee \lceil \frac{l-b}{t} \rceil)q - 1}{I - 1}$$

where $\lceil \cdot \rceil$ is the ceiling function, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$

Based on these results and a first-order approximation $\mathbb{E}(A/B) \approx \mathbb{E}(A)/\mathbb{E}(B)$, we obtained the following formula to approximate the expected value of $\mathrm{Var}(\hat{\theta}|P)$:

$$\mathbb{E}_P(\mathrm{Var}(\hat{\theta}|P)) \approx \frac{fT(f + gT)}{fT(f + gT)\mathbb{E}(\ell - z) - (f + gT)\mathbb{E}(y^2) - f\mathbb{E}(Tw - \ell^2)} \tag{2}$$

where

$$\mathbb{E}(\ell - z) = \frac{T - b + t}{2}\left(\frac{N_{SW}}{\sigma_e^2} - \frac{1}{3T}\left(\frac{N_{SW}}{\sigma_e^2} - f\right)(2T - 2b + t)\right),$$

$$N_{SW} = \sum_{i=1}^{I} n_i,$$

$$\mathbb{E}(y^2) = \frac{T-b+t}{12(I-1)}(s_1 I(T-b-t) + f^2(3I(T-b+t) - 2(2T-2b+t)),$$
$$s_1 = \sum_{i=1}^{I} \frac{1}{(\sigma_i^2 + T\tau^2)^2},$$
$$\mathbb{E}(Tw - \ell^2) = \frac{(T-b+t)N_{SW}^2}{12(T-b)\sigma_e^4}\left(\frac{(T+b)(T-b-t)}{I}\kappa^2 + T^2 + 2bT - tT - 3b^2 + 3bt\right),$$
$$\kappa^2 = \frac{I^2}{N_{SW}^2(I-1)}\sum_{i=1}^{I}\left(n_i - \frac{N_{SW}}{I}\right)^2$$

and here we use κ to denote the sample coefficient of variation (CV) of cluster sizes.

2.5 *Approximate Expected Treatment Effect Variance with Mean and Coefficient of Variation (CV) of Cluster Sizes*

In the design stage of a SW-CRT, each cluster size may not be known, instead, the investigators may have information on the average and CV of the cluster sizes. We derived the following variance formula to approximate the variance formula provided in equation 8 of Hussey and Hughes (2007) for the same type of design described in Section 2.4 where information on actual cluster sizes n_i is replaced by their mean and CV:

$$\mathbb{E}_P(\text{Var}(\hat{\theta}|P)) \approx \frac{IT\sigma^2(\sigma^2 + T\tau^2)}{\sigma^2(ITU - U^2 - I^2C) + T\tau^2(ITU - IV - I^2C)} \tag{3}$$

where

$$C = \frac{(T-b+t)}{12(T-b)}\left(\frac{(T+b)(T-b-t)}{I}\kappa^2 + T^2 + 2bT - tT - 3b^2 + 3bt\right)$$
$$U = \frac{I(T-b+t)}{2}, \quad V = \frac{I(T-b+t)(2T-2b+t)}{6}, \quad \sigma^2 = \frac{I\sigma_e^2}{N_{SW}}$$

The derivation used the first order Taylor approximation of the terms f and s_1 about the mean cluster size (See Web Appendix E for details).

2.6 *Design Effect (DE), Correction Factor (CF) and Relative Efficiency (RE) for Designs with Unequal Cluster Sizes*

From the approximation of the treatment effect variance in Section 2.5, we derived a DE for a SW-CRT with unequal cluster sizes relative to an individually randomized trial of total size $N_{SW} = nI$, the number of participants sampled at each time-point of a cross-sectional SW-CRT (See Web Appendix F):

$$DE_{w,\kappa} = \frac{3(T-b)T(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho))} \tag{4}$$

where $\rho = \tau^2/(\sigma_e^2+\tau^2)$ is the intra-cluster correlation and $n = N_{SW}/I$ is the mean cluster size. When the CV denoted by κ in equation 4 is zero, $DE_{w,\kappa}$ reduces to DE_w a design effect derived in Woertman et al. (2013) for a cross-sectional SW-CRT with equal cluster sizes compared to an individually randomized trial of size N_{SW} (See Web Appendix G). $DE_{w,\kappa}$ summarizes the inflation required for a SW-CRT with unequal cluster sizes compared to an individually randomized trial of total size N_{SW}. To compare to an individually randomized trial of total size TN_{SW}, the overall number of participants sampled in a cross-sectional SW-CRT, the design effect would be $T\ DE_{w,\kappa}$.

In addition, we derive a correction factor (CF) in total sample size at each time-point accounting for varying cluster sizes (See Web Appendix H):

$$N_{SW} \approx DE_w N_{ind} + CF \tag{5}$$

where

- $N_{ind} = 4(\sigma_e^2+\tau^2)(z_{1-\beta}+z_{1-\alpha/2})^2/\theta_A^2$ is the total sample size required for an individually randomized trial with an anticipated treatment effect of θ_A
- $DE_w = [3(T-b)(1-\rho)(1-\rho+T\rho n)]/[(T-b+t)(T-b-t)(2(1-\rho)+(T+b)n\rho)]$ is the Woertman et al. (2013) design effect
- $CF = n\kappa^2\big(1-AT\big)$ is the correction factor for cluster size variation with an attenuation term (AT) defined as $AT = (T-b)(1-\rho)/[T(2(1-\rho)+(T+b)n\rho)]$

The required sample size at each time-point for a SW-CRT can be calculated by multiplying the unadjusted sample size in an individually randomized trial by the Woertman et al. (2013) DE_w, and then adding our CF. Before applying the sample size formula knowledge of the following is required: the mean cluster size (n), the cluster size CV (κ), the number of baseline time-points (b), the number of time-points between each step (t), and the total number of time-points equal to the number of steps (K) multiplied by the number of time-points between each step (t) plus the baseline time-points (b), i.e. $T = Kt + b$. The formulas for individually randomized trials usually result in the number of participants per treatment arm, but the total number of participants is needed here. After utilizing the formula, the required number of clusters I is $\lceil N_{SW}/n \rceil$ and the number of clusters switching treatment at each step is $\lceil I/K \rceil$, where $\lceil \cdot \rceil$ is the ceiling function. Woertman et al. (2013) suggest distributing the clusters as evenly as possible over the steps. As noted in Baio et al. (2015), the overall sample size in terms of participants each contributing one measurement in a cross-sectional design is actually $TN_{SW} = T\ DE_w\ N_{ind}$ if the cluster size is considered fixed and $TN_{SW} = T(DE_w\ N_{ind} + CF)$ using our formulation that accounts for cluster size variation.

The correction factor CF slightly underestimates $n\kappa^2$ and often times can simply be approximated by $n\kappa^2$, since the attenuation term AT is typically negligible. When n is large (e.g. 5,000), there is a high intra-cluster correlation (e.g. $\rho = 0.4$), there are a few baseline time-points (e.g. $b = 2$), and a small number of total time-points (e.g. $T = 4$), we can calculate $AT = 6.25 \times 10^{-6}$. In a more extreme small-sample scenario, when n is small (e.g. 5), intra-cluster correlation is low (e.g. $\rho = 0.01$), there is one baseline time-point (e.g. $b = 1$), and the total time-points is large (e.g. $T = 25$), we can calculate $AT = 0.29$.

In many cases, the CF corresponds to including κ^2 additional clusters to account for cluster size variation. The calculation requires the total number of time-points (T) to remain constant, and hence the number of steps (K) to remain the same. Therefore, it would only

be directly applicable if κ^2 is divisible by the number of steps (K) in the proposed design. For example, with a CV of approximately 1.4 ($\kappa \approx 1.4$) and a design with two steps ($K = 2$), you would include two additional clusters, one per step, in the design to account for cluster size variation.

The approximate expected Relative Efficiency (RE) of a design with unequal cluster sizes compared to equal can be derived (See Web Appendix I) by the ratio of the DE_w derived by Woertman et al. (2013) to our $DE_{w,\kappa}$ in equation 4:

$$RE \approx \frac{DE_w}{DE_{w,\kappa}} = 1 - \frac{\kappa^2}{I}\Big(1 - AT\Big) \tag{6}$$

As noted above, the attenuation term AT is positive and typically small, so the efficiency loss by having unequal cluster sizes can be approximated as $1 - \kappa^2/I$ in most cases.

3. Simulation Study

3.1 *Simulation Study Design*

In our proposed method, cluster i has size n_i, which is fixed for all sampling time-points T. At each of the T time-points a total of N_{SW} participants contribute data from the I clusters. For each simulation scenario, the total number of participants contributing data (N_{SW}) at each time-point was kept fixed and so was the number of participants contributing data from each cluster, i.e. n_i for $i = 1, ..., I$. The number of participants contributing data from each cluster (n_i for $i = 1, ..., I$) was determined by the following procedure. Firstly, the total number of participants contributing data at each time-point (N_{SW}) was randomly split into two groups, with one group containing on average 50%, 60%, 70%, 80% or 90% of the participants, then either: 1. within each group, participants were randomly assigned to one of $I/2$ clusters with equal chance, 2. all the participants from the smaller group were assigned to one cluster, and participants in the larger group were randomly assigned to the remaining $I - 1$ clusters with equal chance, or 3. all the participants from the larger group

were assigned to one cluster, and participants in the smaller group were randomly assigned to the remaining $I-1$ clusters with equal chance. This procedure created cluster size imbalance so that the CV of cluster size variation (κ) ranged from 0 to a maximum of 3.5.

For each chosen N_{SW} and study design, the effect size (θ_A) was set so that the power calculated based on a fixed cluster size using the variance in equation 8 of Hussey and Hughes (2007) would be 80%. Then the estimated power accounting for cluster size variation when all cluster sizes are known was calculated using the variance formula in equation 2. Additionally, the approximate power when only a cluster size mean (n) and CV (κ) is known prior to randomization using equation 3 was calculated. To estimate the empirical power, for each SW-CRT the variable size clusters were placed in a random order and data were simulated using the model given in Section 2.1 (Hussey and Hughes, 2007). For convenience, both μ and β_j for $j = 1, .., T-1$ were set at zero. For continuous outcomes, without loss of generality, the total variance $\sigma_t^2 = \sigma_e^2 + \tau^2$ was fixed at 1, so that the between-cluster and within-cluster variances could then be written as $\tau^2 = \rho$ and $\sigma_e^2 = 1-\rho$, respectively, where ρ is the intra-cluster correlation. Data were analyzed by the same linear mixed effect model using the "lmer" function from the "lme4" package in R statistical software. For simulations involving count outcomes, $\sigma^2 = (1+e^{\theta_A})/2$ and $\tau^2 = (\rho\sigma^2)/(1-\rho)$. α_i were drawn from independent $N(0, \tau^2)$ and $exp(\alpha_i + X_{ij}\theta_A)$ calculated. Count data were then derived from a Poisson distribution with rate $exp(\alpha_i + X_{ij}\theta_A)$. Data were analyzed using the generalized linear mixed effects model with log link, implemented in the R function "glmer". For all simulations, a two-tailed Wald test for the treatment effect was generated, and the empirical power calculated by the proportion of simulated results with p-values that were < 0.05.

In the simulations presented, the intra-cluster correlation ρ was set at 0.05 and the mean cluster size n at 30. For continuous outcomes, we simulated the cases where there were 4 or 6 clusters ($I = 4$ or 6), $q = 1$ cluster randomized at each step, $b = 1$ baseline time-point and

$t = 1$ time-point between each step. Additionally, we simulated the case involving 12 clusters ($I = 12$) where $q = 3$ were randomized at each step, there were $b = 2$ baseline time-points and $t = 3$ time-points between each step. For count data we used the case with $I = 6$, $q = 1$, $b = 1$ and $t = 1$. A Monte Carlo estimate of the error around the empirical power was computed for two cases (first: I=4, q=1, t=1, b=1, n=30, CV=0.73; second: I=12, q=3, t=3, b=2, n=30, CV=0.69) for continuous outcomes and used to guide the choice of the number of simulations. Using 3,500 simulations resulted in a Monte Carlo error of $\leqslant 0.75\%$, so we would expect most empirical power estimates to be within 1.5% of the true power. In the case with $I = 4$, $q = 1$, $b = 1$ and $t = 1$, there would be a total of 120 participants measured at each cross-sectional time-point ($N_{SW} = 120$) and 600 participants contributing data over the course of SW-CRT ($TN_{SW} = 600$).

To evaluate the impact of the order of randomization for particular known cluster sizes for a design with 6 clusters ($I = 6$) and a continuous outcome, the power was calculated for each of the $6! = 720$ randomization sequences using equation 1. We then used the method described in Section 2.3 to create upper and lower bounds for the power across all randomization sequences, and additionally simulated the empirical power for each randomization sequence to observe how close the estimated power using equation 1 was to the truth.

3.2 *Simulation Study Results*

Figure 1 displays the power as the cluster size varies. For the four simulation scenarios one can observe that as the cluster size CV (κ) increases the power decreases. The expected power when all the clusters sizes are known calculated using the variance formula in equation 2, displayed by the dotted light blue line, is very similar to the approximation using the cluster size mean and CV by equation 3 denoted by the solid pink line. This figure appears in color in the electronic version of this article, and color refers to that version. By comparing the orange squares showing the empirical power, we can see the expected power is well estimated.

Most of the deviation of the orange squares from the dotted light blue and solid pink lines is due to the particular order clusters received intervention in that simulation as we know the Monte Carlo error is generally $\leqslant 0.75\%$. The impact of the randomization sequence is explored in more detail in the next paragraph.

[Figure 1 about here.]

Figure 2 displays the power for each possible randomization sequence of 6 clusters where the cluster sizes are 4, 11, 18, 21, 22 and 104, resulting in an average cluster size of 30. If all clusters had equal sample size of 30, the study would have achieved 80% power. Using equation 2 where all clusters sizes are known prior to randomization or equation 3 when only the cluster size mean and CV of $\kappa = 1.23$ is known in the design stage, the study would have an average power of just under 70%. Power estimates for each randomization sequence using equation 1 are displayed by the green dots in Figure 2A. This figure appears in color in the electronic version of this article, and color refers to that version. The green dots are almost entirely covered by the orange squares displaying the empirically simulated power in Figure 2B. An upper and lower bound for the power was obtained using the method described in Section 2.3 and is displayed in Figure 2 by dashed dark blue lines. This correctly finds the maximum and minimum power for estimates derived using equation 1 for the treatment effect variance displayed by green dots. For the empirically simulated power, the lower and upper bounds are very close to the lowest and highest simulated powers.

[Figure 2 about here.]

4. Illustrative Examples

4.1 *Design of a Cross-sectional SW-CRT with Unequal Cluster Sizes*

Suppose we are interested in designing a cross-sectional SW-CRT with two steps ($K = 2$), one baseline time-point ($b = 1$) and one time-point between each step ($t = 1$) to detect

a treatment effect (θ_A) of 0.27 with a mean cluster size (n) of 100. Assuming no cluster size variation, the study would need 6 clusters ($I = 6$) with 3 randomized at each step ($q = 3$). Based on equation 5, under cluster size variation, the overall sample size (TN_{SW}) and number of clusters (I) required to achieve 80% power is plotted as a function of CV (κ) and is shown in Figure 3. If the cluster size CV was $\kappa = 1.4$, the SW-CRT design with two steps would instead require a total of 8 clusters ($I = 8$) with 4 randomized at each step ($q = 4$) (Figure 3B). This represents an inflation of overall sample size (TN_{SW}) from 1800 to 2400, which is 33% (Figure 3A). In this example with only two steps and CV of $\kappa = 1.4$ requiring two additional clusters is easy to implement into the original design, however in many cases this will not be the case. If the CV was 1, only one additional cluster would be required. It is unknown if this cluster should be randomized to intervention at the first or second step. Therefore, the design may have to be changed to have a different number of steps or larger mean cluster size, and finding a design that is practical with adequate power may be an iterative process.

[Figure 3 about here.]

4.2 *Relative Efficiency (RE) Comparing Unequal to Equal Cluster Sizes*

We take as the first example a case where there are a small number of clusters ($I = 4$), so that the RE of the SW-CRT design is notably reduced by cluster size variation. Suppose that one cluster is randomized at each step ($q = 1$), there is one baseline time-point ($b = 1$), one time-point between each step ($t = 1$), the mean cluster size (n) is either 30 or 100, the intra-cluster correlation (ρ) is 0.01, 0.05 or 0.25, and the CV (κ) of cluster size ranges from zero to 1.5. Then the RE for the SW-CRT with unequal cluster sizes to equal is displayed in Figure 4A. With a CV of $\kappa = 0.5$, there is roughly 5% efficiency loss and for a CV of $\kappa = 0.75$ there is greater than 10% efficiency loss in all cases. The efficiency loss is similar for a mean cluster size (n) of 30 or 100 when the intra-cluster correlation is above 0.01 ($\rho > 0.01$).

For the second example, we illustrate that how you divide the clusters over steps does not have much impact on the RE for unequal compared to equal cluster sizes. We take the number of clusters (I) to be 12, the mean cluster size (n) to be either 30 or 100, the intra-cluster correlation (ρ) to be 0.05, the number of clusters randomized at each step (q) to be 1, 2, 3 or 4, and CV (κ) to range from zero to 1.5. On Figure 4B, note that because there are more clusters than in the first example, the impact of cluster size variation overall is less. However, when the CV (κ) is greater than around 1.1 there is more than 10% efficiency loss. The efficiency loss is almost identical whether the average cluster size is 30 or 100, or if a different number of clusters are randomized at each step.

[Figure 4 about here.]

5. Discussion

When designing a cross-sectional SW-CRT with varying cluster sizes, this variation needs to be taken into consideration to ensure the study is adequately powered. While the effect of unequal cluster sizes on study power appears to be smaller than for standard CRTs; the reduction in power is not negligible particularly when the number of clusters is small or the cluster size CV is greater than one.

We derived analytical formulas for calculating the power of a cross-sectional SW-CRT accounting for cluster size variation. In the presence of unequal cluster sizes, the power of a SW-CRT depends on the order of randomization. The variance formula derived in equation 1 is associated with a particular realization of the randomization. Based on this, we have devised computationally efficient algorithms to identify the upper and lower bounds for power without having to resort to an exhaustive search.

Under settings where an equal number of clusters are randomized to treatment initiation at each step, we provide formulas to estimate the average power loss using either the actual

cluster sizes or the projected average and CV of the cluster sizes and note that the power loss can be substantial. If the number of clusters are sufficiently large so that several clusters are randomized to treatment initiation and there is no substantial inequality in the total size of all clusters randomized to treatment initiation at each step, the power loss will be alleviated (Girling, 2018). However, in settings where total numbers of clusters are not large, this would not be feasible. Indeed, SW-CRTs are particularly suited for the situation where the number of clusters is small.

The linear mixed effects model used in this paper, based on work by Hussey and Hughes (2007), assumes random cluster effects, fixed time effects, no cluster by time interaction and no treatment by time interaction. In theory, the model for cross-sectional SW-CRT designs could be adapted to incorporate more flexible modeling assumptions, and a variance derived in a similar fashion by the WLS approach to estimate the power. Alternatively, for more complicated analyses, a simulation approach similar to that proposed by Baio et al. (2015) may be employed.

Further work is necessary for cohort SW-CRTs. Hooper et al. (2016) proposed DEs for cohort designs with equal cluster sizes. Utilizing the DE for cross-sectional design for a cohort study will often result in an underpowered study as the autocorrelation between time-points is not taken into account. As the average RE of a cross-sectional SW-CRT with unequal compared to equal cluster sizes does depend on the intra-cluster correlation (ρ) we expect in a cohort design the autocorrelation between time-points to play a role in determining the RE of unequal compared to equal cluster sizes for a cohort SW-CRT. However, the efficiency loss for a cross-sectional SW-CRT was largely driven by the number of clusters and cluster size CV. This may also be the case for cohort designs, and further exploration is needed.

The formulas for variance and DE in this paper are derived under a linear model and therefore are particularly suited to a continuous or count outcome. For a SW-CRT with a

binary outcome, Zhou et al. (2018) pointed out that the power formulas derived under a linear model can be liberal in some settings and conservative under others, and proposed a method for power calculation based on a likelihood approach. Li et al. (2018) proposed a method to determine sample size for binary outcomes within the framework of generalized estimating equations. Both methods assume equal cluster sizes. In future work we aspire to investigate power and sample size formulas for binary outcomes accounting for unequal cluster sizes.

Acknowledgements

We gratefully acknowledge grants from National Institute of Allergy and Infectious Disease T32 AI 007358, R37 AI 051164 and R01 AI136947. We thank Xuan Zhang for helpful discussions on devising efficient algorithms for identification of bounds.

Supplementary Materials

Web Appendix A and B referenced in Section 2.2, Web Appendix C referenced in Section 2.3, Web Appendix D referenced in Section 2.4, Web Appendix E referenced in Section 2.5, and Web Appendix F, G, H and I referenced in Section 2.6 are available with this paper at the Biometrics website on Wiley Online Library.

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,** 354.

Bashour, H. N., Kanaan, M., Kharouf, M. H., Abdulsalam, A. A., Tabbaa, M. A., and Cheikha, S. A. (2013). The effect of training doctors in communication skills on women's satisfaction with doctor-woman relationship during labour and delivery: a stepped wedge cluster randomised trial in damascus. *BMJ Open* **3,**.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* **6,** 54.

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

Eldridge, S. M., Ashby, D., and Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* **35,** 1292–300.

Girling, A. J. (2018). Relative efficiency of unequal cluster sizes in stepped wedge and other trial designs under longitudinal or cross-sectional sampling. *Stat Med* .

Gurobi (2018). Gurobi optimizer. http://www.gurobi.com.

Hayes, R. and Moulton, L. (2009). *Cluster Randomised Trials*. CRC Press.

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,** h391.

Hill, A. M., McPhail, S. M., Waldron, N., Etherton-Beer, C., Ingram, K., Flicker, L., Bulsara, M., and Haines, T. P. (2015). Fall rates in hospital rehabilitation units after individualised patient and staff education programmes: a pragmatic, stepped-wedge, cluster-randomised controlled trial. *Lancet* **385,** 2592–9.

Hill, A. M., Waldron, N., Etherton-Beer, C., McPhail, S. M., Ingram, K., Flicker, L., and Haines, T. P. (2014). A stepped-wedge cluster randomised controlled trial for evaluating rates of falls among inpatients in aged care rehabilitation units receiving tailored multimedia education in addition to usual care: a trial protocol. *BMJ Open* **4,** e004195.

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

Hoover, D. R. (2002). Power for t-test comparisons of unbalanced cluster exposure studies. *J Urban Health* **79,** 278–94.

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* **28,** 182–91.

Kerry, S. M. and Bland, J. M. (2001). Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Stat Med* **20,** 377–90.

Kristunas, C., Morris, T., and Gray, L. (2017). Unequal cluster sizes in stepped-wedge cluster randomised trials: a systematic review. *BMJ Open* **7,** e017151.

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,** 109.

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

Manatunga, A. K., Hudgens, M. G., and Chen, S. (2001). Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* **43,** 75–86.

Martin, J., Hemming, K., and Girling, A. (2018). The impact of varying cluster size on precision in cross-sectional stepped-wedge cluster randomised trials. In *Second International Conference on Stepped-Wedge Trial Design*.

Matthews, J. N. (2016). The design of stepped wedge trials with unequal cluster sizes. In *First International Conference on Stepped Wedge Trial Design*, volume 17 Suppl 1, page 311. Trials.

Poldervaart, J. M., Reitsma, J. B., Backus, B. E., Koffijberg, H., Veldkamp, R. F., Ten Haaf, M. E., Appelman, Y., Mannaerts, H. F. J., van Dantzig, J. M., van den Heuvel, M., El Farissi, M., Rensing, B., Ernst, N., Dekker, I. M. C., den Hartog, F. R., Oosterhof, T., Lagerweij, G. R., and Buijs, E. M. (2017). Effect of using the heart score in patients

with chest pain in the emergency department: A stepped-wedge, cluster randomized trial. *Ann Intern Med* **166,** 689–697.

Poldervaart, J. M., Reitsma, J. B., Koffijberg, H., Backus, B. E., Six, A. J., Doevendans, P. A., and Hoes, A. W. (2013). The impact of the heart risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. *BMC Cardiovasc Disord* **13,** 77.

van Breukelen, G. J., Candel, M. J., and Berger, M. P. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med* **26,** 2589–603.

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. *J Clin Epidemiol* **66,** 752–8.

Zhou, X., Liao, X., Kunz, L. M., Normand, S. T., Wang, M., and Spiegelman, D. (2018). A maximum likelihood approach to power calculations for stepped wedge designs of binary outcomes. *Biostatistics* .

Received December 2018. *Revised January* 2019. *Accepted February* 2019.

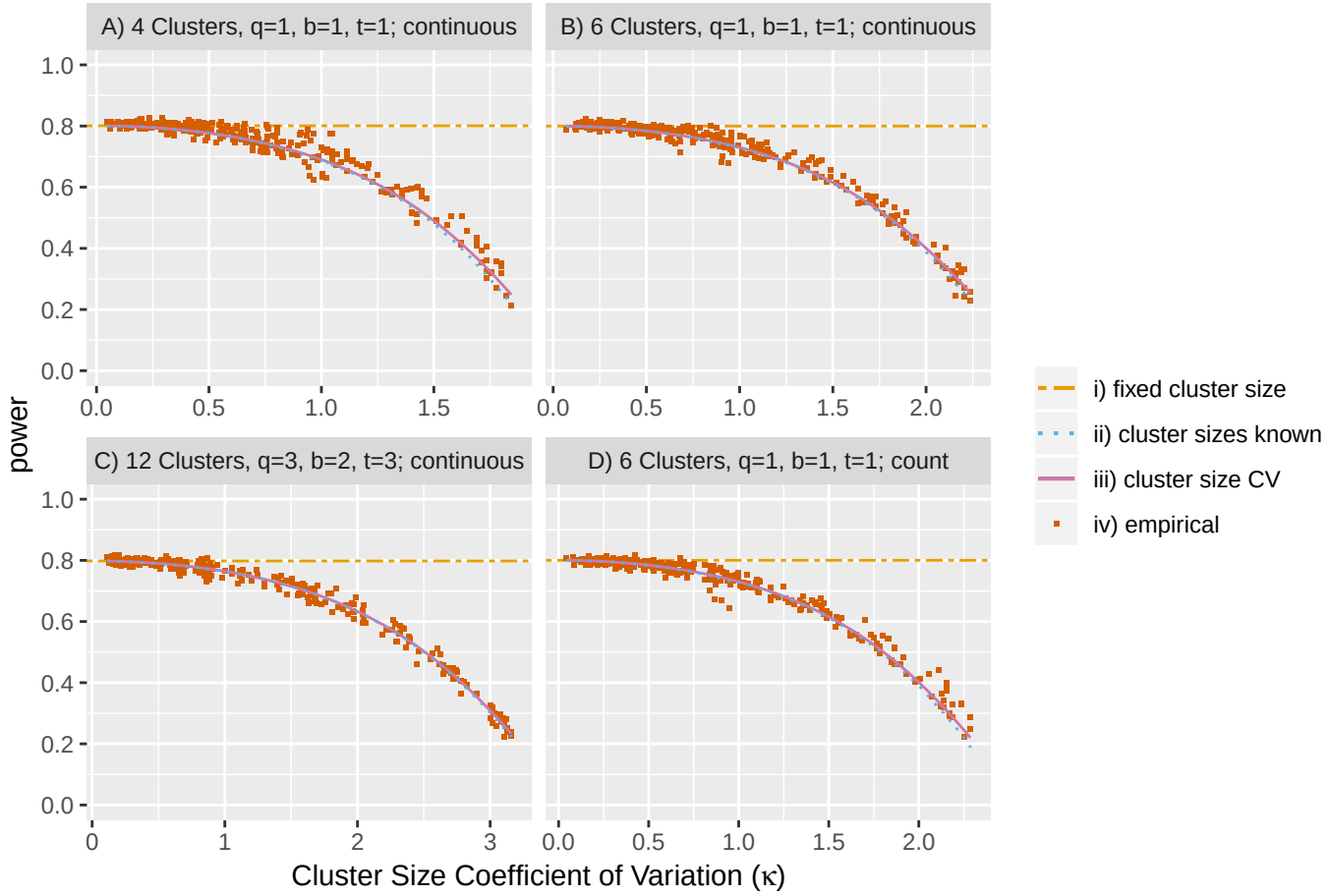

Figure 1. Estimated Power as the Cluster Size Coefficient of Variation (κ) Increases for Four Simulation Scenarios. q is the number of clusters randomized at each step, b is the number of time-points each cluster contributes samples at baseline, t is the number of time-points each cluster contributes samples between each step. i) fixed cluster size: uses the variance formula in equation 8 of Hussey and Hughes (2007), ii) cluster sizes known: uses the variance formula in equation 2, iii) cluster size CV: uses the variance formula in equation 3, iv) empirical: the empirically simulated power from 3,500 simulations. This figure appears in color in the electronic version of this article.

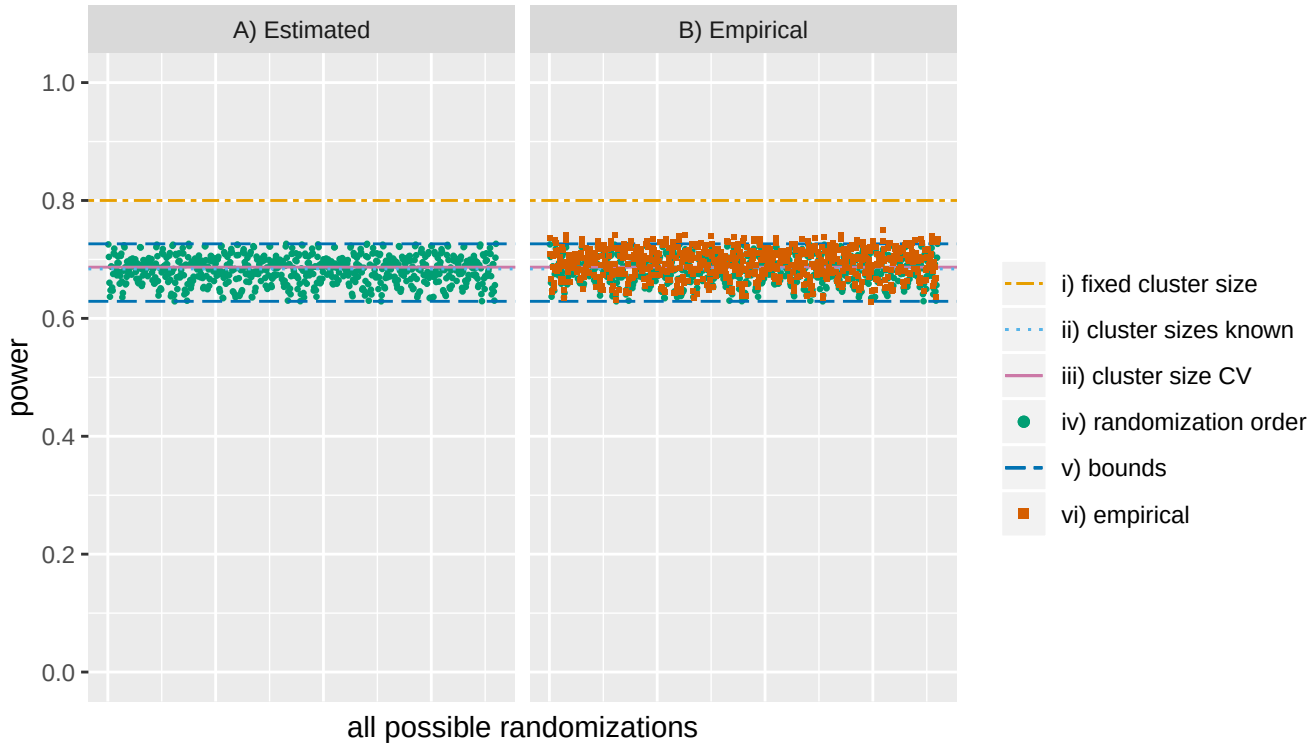

Figure 2. Power for All Possible Randomizations with Upper and Lower Bound Indicated for a SW-CRT with 6 Clusters of Mean Size (n) 30 and CV (κ) of 1.23. i) fixed cluster size: uses the variance formula in equation 8 of Hussey and Hughes (2007), ii) cluster sizes known: uses the variance formula in equation 2, iii) cluster size CV: uses the variance formula in equation 3, iv) randomization order: uses the variance formula in equation 1, v) bounds: uses the method described in Section 2.3, vi) empirical: the empirically simulated power from 3,500 simulations. This figure appears in color in the electronic version of this article.

Figure 3. A) Overall Sample Size (TN_{SW}), B) Number of Clusters (I) Required to Achieve 80% Power for a SW-CRT with Two Steps. i) fixed cluster size: estimates the sample size using DE_w (Woertman et al., 2013), ii) cluster size CV: estimates the sample size using equation 5. This figure appears in color in the electronic version of this article.

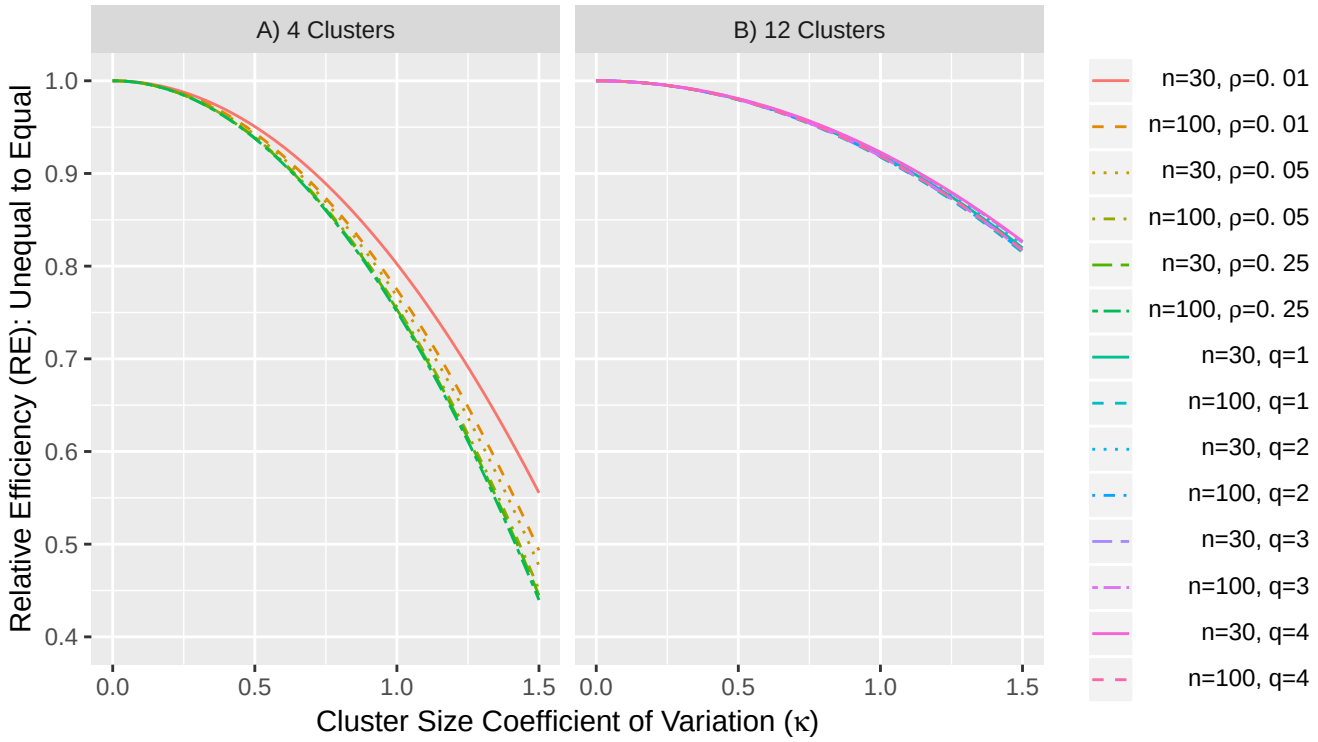

Figure 4. Relative Efficiency (RE) for A) Four Clusters with q=1 Randomized at Each Step (T=5), B) Twelve Clusters with q=1-4 Randomized at Each Step (T=4, 5, 7, 13). n is the mean cluster size and ρ the intra-cluster correlation. This figure appears in color in the electronic version of this article.

Table 1
Schematic of a Stepped Wedge Cluster Randomized Trial (SW-CRT) with one baseline time-point ($b = 1$), two clusters randomized to initiate the intervention at each step ($q = 2$), two time-points between each step ($t = 2$) and four steps ($K = 4$). 0 represents control and 1 intervention.

		Time								
		1	2	3	4	5	6	7	8	9
Cluster	1	0	1	1	1	1	1	1	1	1
	2	0	1	1	1	1	1	1	1	1
	3	0	0	0	1	1	1	1	1	1
	4	0	0	0	1	1	1	1	1	1
	5	0	0	0	0	0	1	1	1	1
	6	0	0	0	0	0	1	1	1	1
	7	0	0	0	0	0	0	0	1	1
	8	0	0	0	0	0	0	0	1	1

Web-based Supplementary Material for Power Calculation for Cross-Sectional Stepped Wedge Cluster Randomized Trials with Variable Cluster Sizes

Linda J Harrison[1,*], Tom Chen[1,2], and Rui Wang[1,2]

[1]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, U.S.A

[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, U.S.A.

Web Appendix A

Let n_i denote the size of each cluster $i = 1, \cdots, I$ at every time-point $j = 1, \cdots, T$. Let $\sigma_i^2 = \sigma_e^2/n_i$. The design matrix $Z \in \mathbb{R}^{IT \times (T+1)}$ takes the form

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_I \end{bmatrix}, \quad \text{where} \quad Z_i = \left[\begin{array}{c|c|c} \mathbf{1}_T & \mathbf{I}_{T-1} & \mathbf{X}_i \\ & \mathbf{0}_{T-1}^\intercal & \end{array}\right]$$

where $\mathbf{1}_T$ and $\mathbf{0}_T$ are vectors of 1's and 0's of length T, respectively, $\mathbf{I}_T$ is the $T \times T$ identity matrix, and $\mathbf{X_i} = (X_{i1}, \cdots, X_{iT})^\intercal$ denotes the treatment status of cluster i at each time j, $j = 1, ..., T$. The inverse of the variance matrix of the outcome vector (Y_{ijk}), $k = 1, ..., n_i$, $V \in \mathbb{R}^{IT \times IT}$ is block-diagonal, $V = \text{diag}(V_1, \cdots, V_I)$, with each block

$$V_i = \sigma_i^2 \mathbf{I}_T + \tau^2 \mathbf{1}_T \mathbf{1}_T^\intercal$$

Therefore, block-matrix multiplication produces

$$Z^\intercal V^{-1} Z = \sum_{i=1}^{I} Z_i^\intercal V_i^{-1} Z_i$$

By Woodbury's formula,

$$V_i^{-1} = \frac{1}{\sigma_i^2(\sigma_i^2 + T\tau^2)}[(\sigma_i^2 + T\tau^2)\mathbf{I}_T - \tau^2 \mathbf{1}_T \mathbf{1}_T^\intercal]$$

And so

$$\begin{aligned} Z_i^\intercal V_i^{-1} Z_i &= \frac{1}{\sigma_i^2(\sigma_i^2 + T\tau^2)}[(\sigma_i^2 + T\tau^2) Z_i^\intercal Z_i - \tau^2 (Z_i^\intercal \mathbf{1})(\mathbf{1}^\intercal Z_i)] \\ &= \frac{1}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \left((\sigma_i^2 + T\tau^2) \begin{bmatrix} T & \mathbf{1}_{T-1}^\intercal & \mathbf{1}_T^\intercal \mathbf{X}_i \\ \mathbf{1}_{T-1} & \mathbf{I}_{T-1} & \mathbf{X}_{i,-T} \\ \mathbf{1}_T^\intercal \mathbf{X}_i & \mathbf{X}_{i,-T}^\intercal & \mathbf{X}_i^\intercal \mathbf{X}_i \end{bmatrix} - \tau^2 \begin{bmatrix} T^2 & T\mathbf{1}_{T-1}^\intercal & T\mathbf{1}_T^\intercal \mathbf{X}_i \\ T\mathbf{1}_{T-1} & \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal & (\mathbf{1}_T^\intercal \mathbf{X}_i)\mathbf{1}_{T-1} \\ T\mathbf{1}_T^\intercal \mathbf{X}_i & (\mathbf{1}_T^\intercal \mathbf{X}_i)\mathbf{1}_{T-1}^\intercal & \mathbf{X}_i^\intercal \mathbf{1}_T \mathbf{1}_T^\intercal \mathbf{X}_i \end{bmatrix} \right) \\ &= \frac{1}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \begin{bmatrix} T\sigma_i^2 & \sigma_i^2 \mathbf{1}_{T-1}^\intercal & \sigma_i^2 \mathbf{1}_T^\intercal \mathbf{X}_i \\ \sigma_i^2 \mathbf{1}_{T-1} & (\sigma_i^2 + T\tau^2)\mathbf{I}_{T-1} - \tau^2 \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal & (\sigma_i^2 + T\tau^2)\mathbf{X}_{i,-T} - \tau^2 (\mathbf{1}_T^\intercal \mathbf{X}_i)\mathbf{1}_{T-1} \\ \sigma_i^2 \mathbf{1}_T^\intercal \mathbf{X}_i & (\sigma_i^2 + T\tau^2)\mathbf{X}_{i,-T}^\intercal - \tau^2 (\mathbf{1}_T^\intercal \mathbf{X}_i)\mathbf{1}_{T-1}^\intercal & (\sigma_i^2 + T\tau^2)\mathbf{X}_i^\intercal \mathbf{X}_i - \tau^2 \mathbf{X}_i^\intercal \mathbf{1}_T \mathbf{1}_T^\intercal \mathbf{X}_i \end{bmatrix} \end{aligned}$$

and so

$$Z^\intercal V^{-1} Z = \begin{bmatrix} Tf & f\mathbf{1}_{T-1}^\intercal & y \\ f\mathbf{1}_{T-1} & (f + gT)\mathbf{I}_{T-1} - g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal & \sum_{i=1}^{I} \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1} \\ y & \sum_{i=1}^{I} \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}^\intercal & \ell - z \end{bmatrix}$$

where

$$f = \sum_{i=1}^{I} \frac{1}{\sigma_i^2 + T\tau^2}, \qquad g = \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)}, \quad \ell = \sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2},$$

$$z = \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \left(\sum_{j=1}^{T} X_{ij} \right)^2, \quad y = \sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2 + T\tau^2}, \quad h = \sum_{i=1}^{I} \sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2(\sigma_i^2 + T\tau^2)}$$

Note the identities

$$f + gT = \sum_{i=1}^{I} \frac{1}{\sigma_i^2}, \qquad \ell = y + T\tau^2 h$$

which we shall freely use in the rest of the proof.

The variance of the treatment effect, $\text{Var}(\hat{\theta})$, is the $(T+1),(T+1)$ component of $(Z^\intercal V^{-1} Z)^{-1}$. Let

$$A_{11} = \begin{bmatrix} Tf & f\mathbf{1}_{T-1}^\intercal \\ f\mathbf{1}_{T-1} & (f+gT)\mathbf{I}_{T-1} - g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal \end{bmatrix}$$

$$A_{21} = A_{12}^\intercal = \begin{bmatrix} y & \sum_{i=1}^I \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}^\intercal \end{bmatrix}$$

$$A_{22} = \ell - z$$

Then,

$$[(Z^\intercal V^{-1} Z)^{-1}]_{(T+1),(T+1)} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

The first task is to compute the components of A_{11}^{-1}, which can be computed through block-matrix inversion as

$$\begin{aligned}
[A_{11}^{-1}]_{11} &= \left((Tf) - f\mathbf{1}_{T-1}^\intercal[(f+gT)\mathbf{I}_{T-1} - g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal]^{-1} f\mathbf{1}_{T-1}\right)^{-1} \\
&= \left((Tf) - f\mathbf{1}_{T-1}^\intercal \frac{(f+g)\mathbf{I}_{T-1} + g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal}{(f+g)(f+gT)} f\mathbf{1}_{T-1}\right)^{-1} \\
&= \left((Tf) - \frac{f^2(f+g)(T-1) + f^2 g(T-1)^2}{(f+g)(f+gT)}\right)^{-1} \\
&= \frac{f+g}{f(f+gT)} \\
[A_{11}^{-1}]_{21} &= [A_{11}^{-1}]_{12}^\intercal = -\left(\frac{(f+g)\mathbf{I}_{T-1} + g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal}{(f+g)(f+gT)}\right)(f\mathbf{1}_{T-1})\left(\frac{f+g}{f(f+gT)}\right) \\
&= -\frac{\mathbf{1}_{T-1}}{f+gT} \\
[A_{11}^{-1}]_{22} &= \left(\frac{(f+g)\mathbf{I}_{T-1} + g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal}{(f+g)(f+gT)}\right) + \frac{\mathbf{1}_{T-1}}{f+gT}(f\mathbf{1}_{T-1}^\intercal)\left(\frac{(f+g)\mathbf{I}_{T-1} + g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal}{(f+g)(f+gT)}\right) \\
&= \frac{1}{f+gT}(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal)
\end{aligned}$$

And so

$$\begin{aligned}
A_{21}A_{11}^{-1}A_{12} &= \begin{bmatrix} y & \sum_{i=1}^I \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}^\intercal \end{bmatrix} \frac{1}{f+gT} \begin{pmatrix} \frac{f+g}{f} & -\mathbf{1}_{T-1}^\intercal \\ -\mathbf{1}_{T-1}^\intercal & \mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal \end{pmatrix} \begin{bmatrix} y \\ \sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1} \end{bmatrix} \\
&= \frac{1}{f+gT}\left(\frac{f+g}{f}y^2 - 2y\eta + \zeta\right)
\end{aligned}$$

where

$$\begin{aligned}
\eta &\stackrel{\text{def}}{=} \mathbf{1}_{T-1}^\intercal \left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}\right) \\
&= \sum_{i=1}^I \sum_{j=1}^{T-1} \frac{X_{ij}}{\sigma_i^2} - \sum_{i=1}^I \sum_{j=1}^T \frac{\tau^2(T-1)X_{ij}}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \\
&= y + \tau^2 h - \sum_{i=1}^I \frac{X_{iT}}{\sigma_i^2}
\end{aligned}$$

and

$$\begin{aligned}
\zeta &\stackrel{\text{def}}{=} \left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}^\intercal\right)(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal)\left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \tau^2 h \mathbf{1}_{T-1}\right) \\
&= \left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \frac{\ell - y}{T}\mathbf{1}_{T-1}^\intercal\right)(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal)\left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \frac{\ell - y}{T}\mathbf{1}_{T-1}\right) \\
&= \left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}^\intercal}{\sigma_i^2} - \frac{\ell}{T}\mathbf{1}_{T-1}^\intercal\right)(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal)\left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \frac{\ell}{T}\mathbf{1}_{T-1}\right) + 2\frac{y}{T}\mathbf{1}_{T-1}^\intercal(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal)\left(\sum_{i=1}^I \frac{\mathbf{X}_{i,-T}}{\sigma_i^2} - \frac{\ell}{T}\mathbf{1}_{T-1}\right)
\end{aligned}$$

$$+ \frac{y}{T}\mathbf{1}_{T-1}^{\intercal}(\mathbf{I}_{T-1} + \mathbf{1}_{T-1}\mathbf{1}_{T-1}^{\intercal})\frac{y}{T}\mathbf{1}_{T-1}$$
$$= \left[w - \frac{\ell^2}{T}\right] + 2y\left(\frac{y}{T} + \tau^2 h - \sum_{i=1}^{I}\frac{X_{iT}}{\sigma_i^2}\right) + \frac{y^2}{T}(T-1)$$

where

$$w = \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{X_{ij}}{\sigma_i^2}\right)^2$$

and so

$$A_{21}A_{11}^{-1}A_{12} = \frac{1}{f+gT}\left(\frac{f+g}{f}y^2 - 2y\left(y + \tau^2 h - \sum_{i=1}^{I}\frac{X_{iT}}{\sigma_i^2}\right) + \left[w - \frac{\ell^2}{T}\right] + 2y\left(\frac{y}{T} + \tau^2 h - \sum_{i=1}^{I}\frac{X_{iT}}{\sigma_i^2}\right) + \frac{y^2}{T}(T-1)\right)$$
$$= \frac{y^2}{fT} + \frac{1}{f+gT}\left(w - \frac{\ell^2}{T}\right)$$

Finally,

$$[(Z^{\intercal}V^{-1}Z)^{-1}]_{(T+1),(T+1)} = \left(\ell - z - \frac{y^2}{fT} - \frac{1}{f+gT}\left(w - \frac{\ell^2}{T}\right)\right)^{-1} = \frac{fT(f+gT)}{fT(f+gT)(\ell - z) - (f+gT)y^2 - f(Tw - \ell^2)}$$

Web Appendix B

When $n_i = n$, then $\sigma_i^2 = \sigma_e^2/n = \sigma^2$ and we can express

$$f = \frac{I}{\sigma^2 + \tau^2 T}, \quad g = \frac{\tau^2 I}{\sigma^2(\sigma^2 + \tau^2 T)}, \quad \ell = \frac{U}{\sigma^2}$$
$$w = \frac{W}{(\sigma^2)^2}, \qquad z = \frac{\tau^2 V}{\sigma^2(\sigma^2 + \tau^2 T)}, \quad y = \frac{U}{\sigma^2 + \tau^2 T}$$

where U, V and W are defined as in equation 8 of Hussey and Hughes (2007). Straightforward algebra yields

$$\mathrm{Var}(\hat{\theta}) = \frac{fT(f+gT)}{fT(f+gT)(\ell - z) - (f+gT)y^2 - f(Tw - \ell^2)}$$
$$= \frac{I\sigma^2(\sigma^2 + \tau^2 T)}{\sigma^2(IU - W) + \tau^2(ITU - IV - TW + U^2)}$$

which is the expression from equation 8 of Hussey and Hughes (2007).

Web Appendix C

Let $\mathbf{v} = (v_1, \cdots, v_I)$ denote a permutation of $\mathbf{n} = (n_1, \cdots, n_I)$. It's understood that ℓ, w, y, z are functions of $\mathbf{v}$, but the dependency is omitted for simplicity. The denominator of the treatment effect variance

$$D(\mathbf{v}) = fT(f+gT)(\ell - z) - (f+gT)y^2 - f(Tw - \ell^2)$$

can be re-expressed in matrix notation as:

$$D(\mathbf{v}) = fT(f+gT)(B^{\intercal}(P\alpha) - (B^2)^{\intercal}(P\beta)) - (f+gT)(P\gamma)^{\intercal}BB^{\intercal}(P\gamma) - f(P\alpha)^{\intercal}(TXX^{\intercal} - BB^{\intercal})(P\alpha)$$

where

$$B^{\intercal} = \left(\sum_{j=1}^{I}X_{1j}, \sum_{j=1}^{I}X_{2j}, ..., \sum_{j=1}^{I}X_{Ij}\right) = (X_{1\cdot}, X_{2\cdot}, ..., X_{I\cdot})$$

$$(B^2)^\intercal = \left(\left(\sum_{j=1}^{I} X_{1j} \right)^2, \left(\sum_{j=1}^{I} X_{2j} \right)^2, ..., \left(\sum_{j=1}^{I} X_{Ij} \right)^2 \right) = (X_{1\cdot}^2, X_{2\cdot}^2, ..., X_{I\cdot}^2)$$

$$\alpha^\intercal = \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_I^2} \right)$$

$$\beta^\intercal = \left(\frac{\tau^2}{\sigma_1^2(\sigma_1^2 + \tau^2 T)}, \frac{\tau^2}{\sigma_2^2(\sigma_2^2 + \tau^2 T)}, \cdots, \frac{\tau^2}{\sigma_I^2(\sigma_I^2 + \tau^2 T)} \right)$$

$$\gamma^\intercal = \left(\frac{1}{\sigma_1^2 + \tau^2 T}, \frac{1}{\sigma_2^2 + \tau^2 T}, \cdots, \frac{1}{\sigma_I^2 + \tau^2 T} \right)$$

where permutation matrix $P \in \{0,1\}^{I \times I}$ satisfies

$$\sum_{i=1}^{I} P_{ij} = 1 \quad \forall j, \qquad \sum_{j=1}^{I} P_{ij} = 1 \quad \forall i$$

Note that the components of α, β, γ are ordered the same as $\mathbf{n}$, and $\mathbf{v} = P\mathbf{n}$ uniquely. Therefore, we may proceed with optimization over permutation matrices P and express the objective as $D(P)$. In order to feed the objective into an optimization package, we need to reformulate the problem into a mixed-integer quadratic programming (MIQP) problem, which requires decision variables in a vector, while our current form is a matrix. Therefore, we vectorize P. To determine the vectorization, we expand the matrix operations:

$$B^\intercal(P\alpha) - (B^2)^\intercal(P\beta) = \sum_{t=1}^{I}\sum_{j=1}^{I} P_{tj} X_{t\cdot}\alpha_j - \sum_{t=1}^{I}\sum_{j=1}^{I} P_{tj} X_{t\cdot}^2 \beta_j = \sum_{t=1}^{I}\sum_{j=1}^{I} P_{tj}(X_{t\cdot}\alpha_j - X_{t\cdot}^2\beta_j)$$

$$(P\gamma)^\intercal BB^\intercal(P\gamma) = \sum_{i=1}^{I}\sum_{j=1}^{I} \gamma_i (P^\intercal BB^\intercal P)_{i,j}\gamma_j = \sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{s=1}^{I}\sum_{t=1}^{I} \gamma_i\gamma_j (BB^\intercal)_{st} P_{si} P_{tj}$$

$$\begin{aligned}(P\alpha)^\intercal(TXX^\intercal - BB^\intercal)(P\alpha) &= \sum_{i=1}^{I}\sum_{j=1}^{I} \alpha_i (P^\intercal(TXX^\intercal - BB^\intercal)P)_{i,j}\alpha_j \\ &= \sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{s=1}^{I}\sum_{t=1}^{I} \alpha_i\alpha_j (TXX^\intercal - BB^\intercal)_{st} P_{si} P_{tj}\end{aligned}$$

The matrix M and vector D is the collection of the coefficients corresponding to the quadratic and linear sums, respectively, which can be simplified into

$$\begin{aligned}M_{(s-1)I+i,(t-1)I+j} = &-(f+gT)\left(\sum_{k=1}^{I} X_{sk}\right)\left(\sum_{k=1}^{I} X_{tk}\right)\frac{1}{(\sigma_i^2 + \tau^2 T)(\sigma_j^2 + \tau^2 T)} \\ &- f\left[T\sum_{k=1}^{I} X_{sk}X_{tk} - \left(\sum_{k=1}^{I} X_{sk}\right)\left(\sum_{k=1}^{I} X_{tk}\right)\right]\frac{1}{\sigma_i^2\sigma_j^2}\end{aligned}$$

and

$$D_{(t-1)I+j} = fT(f+gT)\left[\left(\sum_{k=1}^{I} X_{tk}\right)\frac{1}{\sigma_j^2} - \left(\sum_{k=1}^{I} X_{tk}\right)^2 \frac{\tau^2}{\sigma_j^2(\sigma_j^2 + \tau^2 T)}\right]$$

Web Appendix D

Let's consider a treatment status matrix X, where q clusters are randomized at each step, there are b baseline time-points and t time-points after each step, so the total number of time-points $T = \frac{I}{q}t + b$. Note that

$$\mathbb{P}(X_{i(b+pt+1)} = 1) = \cdots = \mathbb{P}(X_{i(b+(p+1)t)} = 1) \stackrel{\text{def}}{=} \lambda_p$$

for all i, b, t, p. Indeed, for any cluster i, its treatment status at times $\{b+pt+1, \cdots, b+(p+1)t\}$ remain the same; treatment status of a cluster can only change at each step, not at different time-points associated with the same step. That is, $X_{i(b+pt+1)} = \cdots = X_{i(b+(p+1)t)}$. We observe the recursive relation

$$\begin{aligned}\lambda_p &= \mathbb{P}(X_{i(b+(p+1)t)} = 1 | X_{i(b+pt)} = 1)\mathbb{P}(X_{i(b+pt)} = 1) + \mathbb{P}(X_{i(b+(p+1)t)} = 1 | X_{i(b+pt)} = 0)\mathbb{P}(X_{i(b+pt)} = 0) \\ &= 1 \cdot \lambda_{p-1} + \frac{q}{I - pq}(1 - \lambda_{p-1})\end{aligned}$$

with initial condition $\lambda_{-1} = 0$, since no cluster is randomized to treatment initiation before time $b+1$. Through techniques from difference equations (or simply by inspection), we see that

$$\lambda_p = \frac{(p+1)q}{I}$$

solves the recurrence relation. In general, for $k = 1, \cdots, T$,

$$\mathbb{P}(X_{ik} = 1) = \frac{\lceil \frac{k-b}{t} \rceil q}{I}$$

Now let's derive the joint distribution of (X_{ik}, X_{jl}). Assume without loss of generality that $k < l$. Then,

$$\mathbb{P}(X_{ik} = X_{jl} = 1) = \mathbb{P}(X_{ik} = 1)\mathbb{P}(X_{jl} = 1 | X_{ik} = 1) = \frac{\lceil \frac{k-b}{t} \rceil q}{I} \frac{\lceil \frac{l-b}{t} \rceil q - 1}{I - 1}$$

If $k > l$, the variables would change places, hence in general,

$$\mathbb{P}(X_{ik} = X_{jl} = 1) = \frac{(\lceil \frac{k-b}{t} \rceil \wedge \lceil \frac{l-b}{t} \rceil)q}{I} \frac{(\lceil \frac{k-b}{t} \rceil \vee \lceil \frac{l-b}{t} \rceil)q - 1}{I - 1}$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Now we may begin computing expectations. Starting with $\mathbb{E}(\ell)$,

$$\begin{aligned}\mathbb{E}(\ell) = \mathbb{E}\Big(\sum_{i=1}^{I}\sum_{j=1}^{T}\frac{X_{ij}}{\sigma_i^2}\Big) &= \sum_{i=1}^{I}\sum_{p=-1}^{I/q-1}\frac{1}{\sigma_i^2}t\mathbb{E}(X_{i(b+(p+1)t)}) \\ = \sum_{i=1}^{I}\sum_{p=-1}^{I/q-1}\frac{1}{\sigma_i^2}t(p+1)\frac{q}{I} &= \sum_{i=1}^{I}\frac{1}{\sigma_i^2}t\frac{q}{I}\frac{\frac{I}{q}\left(\frac{I}{q}+1\right)}{2} \\ = \sum_{i=1}^{I}\frac{T-b+t}{2\sigma_i^2} &= \frac{T-b+t}{2}(f+gT)\end{aligned}$$

Next for $\mathbb{E}(z)$:

$$\begin{aligned}\mathbb{E}(z) &= \mathbb{E}\left(\sum_{i=1}^{I}\frac{\tau^2}{\sigma_i^2(\sigma_i^2+T\tau^2)}\left(\sum_{j=1}^{T}X_{ij}\right)^2\right) \\ &= \sum_{i=1}^{I}\frac{\tau^2}{\sigma_i^2(\sigma_i^2+T\tau^2)}\mathbb{E}\left(\sum_{j=1}^{T}X_{ij}\right)^2\end{aligned}$$

Expanding the expectation,

$$\begin{aligned}\mathbb{E}\left(\sum_{j=1}^{T}X_{ij}\right)^2 &= \left(\sum_{j=1}^{T}\mathbb{E}(X_{ij}^2)\right) + \left(2\sum_{k<l}\mathbb{E}(X_{ik}X_{il})\right) = \frac{T-b+t}{2} + 2\sum_{k=1}^{T}\frac{\lceil \frac{k-b}{t} \rceil q}{I}(T-k) \\ &= \frac{T-b+t}{2} + 2\sum_{p=-1}^{I/q-1}(p+1)\frac{q}{I}\left(Tt - tb - \frac{t(t+1)}{2} - pt^2\right) \\ &= \frac{T-b+t}{2} + (T-b)(T-b+t) - \frac{(t+1)(T-b+t)}{2} - \frac{(2T-2b-2t)(T-b+t)}{3} \\ &= \frac{(T-b+t)(2T-2b+t)}{6}\end{aligned}$$

And therefore

$$\mathbb{E}(z) = \frac{(T-b+t)(2T-2b+t)}{6}\sum_{i=1}^{I}\frac{\tau^2}{\sigma_i^2(\sigma_i^2+T\tau^2)} = \frac{g(T-b+t)(2T-2b+t)}{6}$$

Note that

$$f + gT = \sum_{i=1}^{I}\frac{1}{\sigma_i^2} = \sum_{i=1}^{I}\frac{n_i}{\sigma_e^2} = \frac{N_{SW}}{\sigma_e^2}$$

where $N_{SW} = \sum_{i=1}^{I} n_i$ and $g = \frac{1}{T}\left(\frac{N_{SW}}{\sigma_e^2} - f\right)$, so,

$$\mathbb{E}(\ell - z) = \frac{T-b+t}{2}\left(\frac{N_{SW}}{\sigma_e^2} - \frac{1}{3T}\left(\frac{N_{SW}}{\sigma_e^2} - f\right)(2T-2b+t)\right)$$

Let $s_1 = \sum_{i=1}^{I}\frac{1}{(\sigma_i^2+T\tau^2)^2}$. Then for $\mathbb{E}(y^2)$:

$$\mathbb{E}(y^2) = \mathbb{E}\left[\left(\sum_{i=1}^{I}\sum_{j=1}^{T}\frac{X_{ij}}{\sigma_i^2+T\tau^2}\right)^2\right] = \frac{(T-b+t)(2T-2b+t)}{6}s_1 + \sum_{i\neq i'}\frac{\sum_{j,j'}\mathbb{E}(X_{ij}X_{i'j'})}{(\sigma_i^2+T\tau^2)(\sigma_{i'}^2+T\tau^2)}$$

We may compute

$$\sum_{j<j'}\mathbb{E}(X_{ij}X_{i'j'}) = \sum_{j=1}^{T}\frac{(\lceil\frac{j-b}{t}\rceil)q}{I}\sum_{j'=j+1}^{T}\frac{(\lceil\frac{j'-b}{t}\rceil)q-1}{I-1}$$

For the innermost sum, we can break the summation range into portions corresponding to (1) $X_{i'j'}$ randomized at the same time point as X_{ij}, for which there are $b + \left\lceil\frac{j-b}{t}\right\rceil t - j$ instances, and (2) $X_{i'j'}$ randomized to a time-point subsequent to X_{ij}, for which there are t instances. Therefore,

$$\begin{aligned}
\sum_{j<j'}\mathbb{E}(X_{ij}X_{i'j'}) &= \sum_{j=1}^{T}\frac{(\lceil\frac{j-b}{t}\rceil)q}{I}\left(\left(b+\left\lceil\frac{j-b}{t}\right\rceil t - j\right)\frac{\lceil\frac{j-b}{t}\rceil q-1}{I-1} + \sum_{p'=\lceil\frac{j-b}{t}\rceil}^{I/q-1}\frac{(p'+1)q-1}{I-1}t\right)\\
&= \sum_{p=-1}^{I/q-1}\frac{(p+1)q}{I}\frac{(p+1)q-1}{I-1}\left(\frac{t(t-1)}{2}\right) + \sum_{p=-1}^{I/q-1}\sum_{p'=p+1}^{I/q-1}\frac{(p+1)q}{I}\frac{(p'+1)q-1}{I-1}t^2\\
&= \frac{(t-1)(T-b+t)(2I+q-3)}{12(I-1)} + \frac{(T-b+t)(T-b-t)(3I+2q-4)}{24(I-1)}
\end{aligned}$$

and

$$\sum_{j=1}^{I}\mathbb{E}(X_{ij}X_{i'j}) = \sum_{p=-1}^{I/q-1}\frac{(p+1)q}{I}\frac{(p+1)q-1}{I-1}t = \frac{(T-b+t)(2I+q-3)}{6(I-1)}$$

so,

$$\begin{aligned}
\sum_{j,j'}\mathbb{E}(X_{ij}X_{i'j'}) &= \frac{(T-b+t)(2I+q-3)}{6(I-1)} + 2\left(\frac{(t-1)(T-b+t)(2I+q-3)}{12(I-1)} + \frac{(T-b+t)(T-b-t)(3I+2q-4)}{24(I-1)}\right)\\
&= \frac{T-b+t}{12(I-1)}[(T-b)(3I-4) + It - 2t + 2qT - 2qb]\\
&= \frac{T-b+t}{12(I-1)}[(T-b)(3I-4) + t(3I-2)]
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}(y^2) &= \frac{(T-b+t)(2T-2b+t)}{6}s_1 + \frac{T-b+t}{12(I-1)}[(T-b)(3I-4)+t(3I-2)](f^2-s_1)\\
&= \left[\frac{(T-b+t)(T-b-t)I}{12(I-1)}\right]s_1 + \left[\frac{(T-b+t)[3I(T-b+t)-2(2T-2b+t)]}{12(I-1)}\right]f^2
\end{aligned}$$

Next,

$$\mathbb{E}(w) = \sum_{j=1}^{T}\left(\sum_{i=1}^{I}\frac{\mathbb{E}(X_{ij}^2)}{\sigma_i^4} + 2\sum_{i<i'}\frac{\mathbb{E}(X_{ij}X_{i'j})}{\sigma_i^2\sigma_{i'}^2}\right) = \frac{T-b+t}{2}\sum_{i=1}^{I}\frac{1}{\sigma_i^4} + \frac{(T-b+t)(2I+q-3)}{6(I-1)}2\sum_{i<i'}\frac{1}{\sigma_i^2\sigma_{i'}^2}$$

$$= \frac{1}{\sigma_e^4}\left(\frac{T-b+t}{2}\left(\sum_{i=1}^{I} n_i^2\right) + \frac{(T-b+t)(2I+q-3)}{6(I-1)}\left(2\sum_{i<i'} n_i n_{i'}\right)\right)$$

and

$$\mathbb{E}(\ell^2) = \frac{(T-b+t)(2T-2b+t)}{6}\sum_{i=1}^{I}\frac{1}{\sigma_i^4} + \frac{(T-b+t)[3I(T-b+t)-2(2T-2b+t)]}{12(I-1)} 2\sum_{i<i'}\frac{1}{\sigma_i^2\sigma_{i'}^2}$$

$$= \frac{1}{\sigma_e^4}\left(\frac{(T-b+t)(2T-2b+t)}{6}\left(\sum_{i=1}^{I} n_i^2\right) + \frac{(T-b+t)[3I(T-b+t)-2(2T-2b+t)]}{12(I-1)}\left(2\sum_{i<i'} n_i n_{i'}\right)\right)$$

with the derivation above following similar steps in the computation of $\mathbb{E}(y^2)$. Hence,

$$\mathbb{E}(Tw-\ell^2) = \frac{T-b+t}{\sigma_e^4}\left(Y_1\left(\sum_{i=1}^{I} n_i^2\right) + Y_2\left(2\sum_{i<i'} n_i n_{i'}\right)\right)$$

where

$$Y_1 = \frac{T+2b-t}{6} \qquad \text{and} \qquad Y_2 = \frac{IT+2qT-2T+3Ib-4b-3tI+2t}{12(I-1)}$$

Let κ denote the sample coefficient of variation (CV) for cluster sizes n_i, we have:

$$\kappa^2 = \frac{\frac{1}{I-1}\sum_{i=1}^{I}\left(n_i - \frac{N_{SW}}{I}\right)^2}{\frac{N_{SW}^2}{I^2}} \iff \sum_{i=1}^{I} n_i^2 = \frac{N_{SW}^2}{I}\left(\frac{I-1}{I}\kappa^2+1\right)$$

we may substitute to yield

$$\begin{aligned}\mathbb{E}(Tw-\ell^2) &= \frac{T-b+t}{\sigma_e^4}\left(\frac{N_{SW}^2}{I}\left(\frac{I-1}{I}\kappa^2+1\right)Y_1 + \left(N_{SW}^2 - \frac{N_{SW}^2}{I}\left(\frac{I-1}{I}\kappa^2+1\right)\right)Y_2\right)\\ &= \frac{(T-b+t)N_{SW}^2}{\sigma_e^4}\left(\frac{(I-1)(Y_1-Y_2)}{I^2}\kappa^2 + Y_2 + \frac{Y_1-Y_2}{I}\right)\\ &= \frac{(T-b+t)N_{SW}^2}{12(T-b)\sigma_e^4}\left(\frac{(T+b)(T-b-t)}{I}\kappa^2 + T^2 + 2bT - tT - 3b^2 + 3bt\right)\end{aligned}$$

Web Appendix E

In order to obtain variance formula similar to equation 8 in Hussey and Hughes (2007) that accounts for cluster size variation, we approximated f and s_1 by their first order Taylor expansion about the mean cluster size:

$$f = \sum_{i=1}^{I}\frac{1}{\sigma_i^2+T\tau^2} = \sum_{i=1}^{I}\frac{n_i}{\sigma_e^2+n_iT\tau^2} \approx \sum_{i=1}^{I}\left[\frac{n}{\sigma_e^2+(n)T\tau^2} + \frac{\sigma_e^2(n_i-n)}{(\sigma_e^2+(n)T\tau^2)^2}\right] = \frac{I}{\frac{\sigma_e^2}{n}+T\tau^2}$$

$$s_1 = \sum_{i=1}^{I}\frac{1}{(\sigma_i^2+T\tau^2)^2} = \sum_{i=1}^{I}\frac{n_i^2}{(\sigma_e^2+n_iT\tau^2)^2} \approx \sum_{i=1}^{I}\left[\frac{(n)^2}{(\sigma_e^2+(n)T\tau^2)^2} + \frac{2(n)\sigma_e^2(n_i-n)}{(\sigma_e^2+(n)T\tau^2)^3}\right] = \frac{I}{\left(\frac{\sigma_e^2}{n}+T\tau^2\right)^2}$$

where n is the mean cluster size. Note that this approximation is exact if $n_i = n$ for all i; that is, cluster sizes do not change. Substituting these approximations,

$$\mathbb{E}(\ell - z) \approx \frac{U}{\sigma^2} - \frac{\tau^2 V}{\sigma^2(\sigma^2+\tau^2T)} \qquad \text{and} \qquad \mathbb{E}(y^2) \approx \left(\frac{U}{\sigma^2+\tau^2T}\right)^2$$

where

$$U = \frac{I(T-b+t)}{2} \qquad \text{and} \qquad V = \frac{I(T-b+t)(2T-2b+t)}{6}$$

We retain

$$\mathbb{E}(Tw-\ell^2) = \frac{(T-b+t)N_{SW}^2}{12(T-b)\sigma_e^4}\left(\frac{(T+b)(T-b-t)}{I}\kappa^2 + T^2 + 2bT - tT - 3b^2 + 3bt\right)$$

to account for cluster size variation.

$$\mathbb{E}_P[\text{Var}(\hat{\theta}|P)] \approx \frac{fT(f+gT)}{fT(f+gT)\mathbb{E}(\ell-z)-(f+gT)\mathbb{E}(y^2)-f\mathbb{E}(Tw-\ell^2)}$$
$$\approx \frac{IT\sigma^2(\sigma^2+T\tau^2)}{\sigma^2(ITU-U^2-I^2C)+T\tau^2(ITU-IV-I^2C)}$$

where

$$C=\frac{(T-b+t)}{12(T-b)}\left(\frac{(T+b)(T-b-t)}{I}\kappa^2+T^2+2bT-tT-3b^2+3bt\right)$$

Web Appendix F

Substituting, $U=I(T-b+t)/2$ and $V=I(T-b+t)(2T-2b+t)/6$ into the approximation from Appendix E,

$$\mathbb{E}_P[\text{Var}(\hat{\theta}|P)] \approx \frac{IT\sigma^2(\sigma^2+T\tau^2)}{\sigma^2\left(IT\frac{I(T-b+t)}{2}-\left(\frac{I(T-b+t)}{2}\right)^2-I^2C\right)+T\tau^2\left(IT\frac{I(T-b+t)}{2}-I\frac{I(T-b+t)(2T-2b+t)}{6}-I^2C\right)}$$
$$=\frac{12(T-b)T\sigma^2(\sigma^2+T\tau^2)}{I(T-b+t)(T-b-t)(\sigma^2(2T-\frac{T+b}{I}\kappa^2)+T\tau^2(T+b-\frac{T+b}{I}\kappa^2))}$$

Let

$$\rho=\frac{\tau^2}{\sigma_e^2+\tau^2}, \sigma_t^2=\sigma_e^2+\tau^2, n=\frac{N_{SW}}{I} \implies \tau^2=\rho\sigma_t^2, \sigma_e^2=\sigma_t^2(1-\rho), \sigma^2=\frac{\sigma_e^2}{n}=\frac{\sigma_t^2(1-\rho)}{n}$$

And so

$$\mathbb{E}_P[\text{Var}(\hat{\theta}|P)] \approx \frac{12(T-b)T\frac{\sigma_t^2(1-\rho)}{n}\left(\frac{\sigma_t^2(1-\rho)}{n}+T\rho\sigma_t^2\right)}{I(T-b+t)(T-b-t)\left(\frac{\sigma_t^2(1-\rho)}{n}(2T-\frac{T+b}{I}\kappa^2)+T\rho\sigma_t^2(T+b-\frac{T+b}{I}\kappa^2)\right)}$$
$$=\frac{12(T-b)T\sigma_t^2(1-\rho)(1-\rho+T\rho n)}{In(T-b+t)(T-b-t)(T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho))}$$

In an individually randomized trial with total sample size $N_{SW}=nI$ and two equally sized treatment groups of size $nI/2$, the T-statistic under the alternative θ_A is $T=\frac{\theta_A}{\sqrt{2\sigma_t^2/(nI/2)}}=\frac{\theta_A}{\sqrt{4\sigma_t^2/(nI)}}$ with the variance of the treatment effect $4\sigma_t^2/(nI)$. Therefore, the design effect for a cross-sectional SW-CRT with unequal cluster sizes is:

$$DE_{w,\kappa}=\frac{\mathbb{E}_P[\text{Var}(\hat{\theta}|P)]}{4\sigma_t^2/(nI)} \approx \frac{3(T-b)T(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho))}$$

Web Appendix G

When $\kappa=0$, then $DE_{w,\kappa}$ reduces to

$$DE_w=\frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(2(1-\rho)+(T+b)n\rho)}$$

The design effect provided in Woertman et al. (2013) is

$$DE_w = \frac{[1+\rho(Ktn+bn-1)]}{[1+\rho(\frac{1}{2}Ktn+bn-1)]}\frac{3(1-\rho)}{2t\Big(K-\frac{1}{K}\Big)}$$

where $K=(T-b)/t$ is the number of steps. Indeed,

$$\begin{aligned}\frac{[1+\rho(Ktn+bn-1)]}{[1+\rho(\frac{1}{2}Ktn+bn-1)]}\frac{3(1-\rho)}{2t\Big(K-\frac{1}{K}\Big)} &= \frac{(1+\rho(nT-1))3(1-\rho)}{(1+\rho(\frac{n}{2}(T+b)-1))2t\Big(\frac{T-b}{t}-\frac{t}{T-b}\Big)}\\ &= \frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(2(1-\rho)+n\rho(T+b))((T-b)(T-b)-t^2)}\\ &= \frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(2(1-\rho)+n\rho(T+b))}\end{aligned}$$

Web Appendix H

The power is

$$1-\beta \approx \Phi\left(\frac{\theta_A}{\sqrt{\mathbb{E}_P[\mathrm{Var}(\hat{\theta}|P)]}}-z_{1-\alpha/2}\right)$$

$$\implies \frac{\theta_A^2}{(z_{1-\beta}+z_{1-\alpha/2})^2} \approx \mathbb{E}_P[\mathrm{Var}(\hat{\theta}|P)] \approx \frac{12(T-b)T\sigma_t^2(1-\rho)(1-\rho+T\rho n)}{In(T-b+t)(T-b-t)(T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho))}$$

Solving for $N_{SW} \stackrel{\text{def}}{=} In$ yields

$$\begin{aligned}N_{SW} &= \frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(2(1-\rho)+(T+b)n\rho)}\frac{4\sigma_t^2(z_{1-\beta}+z_{1-\alpha/2})^2}{\theta_A^2}+\frac{n(T+b)(1-\rho+Tn\rho)\kappa^2}{T(2(1-\rho)+(T+b)n\rho)}\\ &= \frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(2(1-\rho)+(T+b)n\rho)}\frac{4\sigma_t^2(z_{1-\beta}+z_{1-\alpha/2})^2}{\theta_A^2}+n\kappa^2\left(1-\frac{(T-b)(1-\rho)}{T(2(1-\rho)+(T+b)n\rho)}\right)\\ &= DE_w N_{ind}+CF\end{aligned}$$

where DE_w is the DE by Woertman et al. (2013), N_{ind} is the sample size for an individually randomized trial, and CF is the correction factor for cluster size variation.

Web Appendix I

$$\begin{aligned}RE &\approx \frac{DE_w}{DE_{w,\kappa}}\\ &= \frac{\frac{3(T-b)(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(2(1-\rho)+(T+b)n\rho)}}{\frac{3(T-b)T(1-\rho)(1-\rho+T\rho n)}{(T-b+t)(T-b-t)(T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho))}}\\ &= \frac{T(2(1-\rho)+(T+b)n\rho)-\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho)}{T(2(1-\rho)+(T+b)n\rho)}\\ &= 1-\frac{\frac{T+b}{I}\kappa^2(1-\rho+Tn\rho)}{T(2(1-\rho)+(T+b)n\rho)}\end{aligned}$$

$$= 1 - \frac{\kappa^2}{I}\left(1 - \frac{(T-b)(1-\rho)}{T(2(1-\rho)+(T+b)n\rho)}\right)$$

References

Hussey, M. A. and Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* **28,** 182–91.

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. *J Clin Epidemiol* **66,** 752–8.