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

by Linda J Harrison[1,*], Tom Chen[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 $\mathbf{Z} \in \mathbb{R}^{IT \times (T+1)}$ takes the form

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_I \end{bmatrix}, \quad \text{where} \quad \mathbf{Z}_i = \left[\begin{array}{c|c|c} \mathbf{1}_T & \mathbf{I}_{T-1} & \mathbf{X}_i \\ & \mathbf{0}_{T-1}^\intercal & \end{array}\right]$$

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 $j$, $j = 1, ..., T$. Let $\mathbf{V} \in \mathbb{R}^{IT \times IT}$ be the variance matrix of the cluster-level mean responses given by $\frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ijk}$. $\mathbf{V}$ is block-diagonal, $\mathbf{V} = \text{diag}(\mathbf{V}_1, \cdots, \mathbf{V}_I)$, with each block

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

Therefore, block-matrix multiplication produces

$$\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z} = \sum_{i=1}^{I} \mathbf{Z}_i^\intercal \mathbf{V}_i^{-1} \mathbf{Z}_i$$

By Woodbury's formula,

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

And so

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

and so

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

where

$$\begin{aligned}
f &= \sum_{i=1}^{I} \frac{1}{\sigma_i^2 + T\tau^2}, & g &= \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)}, & \ell &= \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2}, \\
z &= \sum_{i=1}^{I} \frac{\tau^2}{\sigma_i^2(\sigma_i^2 + T\tau^2)} \left(\sum_{j=1}^{T} X_{ij}\right)^2, & y &= \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2 + T\tau^2}, & h &= \sum_{i=1}^{I}\sum_{j=1}^{T} \frac{X_{ij}}{\sigma_i^2(\sigma_i^2 + T\tau^2)}
\end{aligned}$$

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, $\mathrm{Var}(\widehat{\theta})$, is the $(T+1),(T+1)$ component of $(\mathbf{Z}^\intercal\mathbf{V}^{-1}\mathbf{Z})^{-1}$. Let

$$\mathbf{A}_{11} = \begin{bmatrix} Tf & f\mathbf{1}_{T-1}^\intercal \\ f\mathbf{1}_{T-1} & (f+gT)\mathbf{I}_{T-1} - g\mathbf{1}_{T-1}\mathbf{1}_{T-1}^\intercal \end{bmatrix}$$
$$\mathbf{A}_{21} = \mathbf{A}_{12}^\intercal = \begin{bmatrix} y & \sum_{i=1}^I \frac{\mathbf{x}_{i,-T}^\intercal}{\sigma_i^2} - \tau^2 h\mathbf{1}_{T-1}^\intercal \end{bmatrix}$$
$$A_{22} = \ell - z$$

Then,

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

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

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

And so

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

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

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

Finally,

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

## 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

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

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 $\ell, w, y, z$ 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:

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

where

$\mathbf{X}$ is the treatment status matrix

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

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

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

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

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

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

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

Note that the components of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are ordered the same as $\mathbf{n}$, and $\mathbf{v} = \mathbf{P}\mathbf{n}$ uniquely. Therefore, we may proceed with optimization over permutation matrices $\mathbf{P}$ and express the objective as $D(\mathbf{P})$. 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 $\mathbf{P}$. To determine the vectorization, we expand the matrix operations:

$$\mathbf{B}^\intercal(\mathbf{P}\boldsymbol{\alpha}) - (\mathbf{B}^2)^\intercal(\mathbf{P}\boldsymbol{\beta}) = \sum_{t=1}^{I}\sum_{j=1}^{I} \mathbf{P}_{tj} X_{t\cdot}\boldsymbol{\alpha}_j - \sum_{t=1}^{I}\sum_{j=1}^{I} \mathbf{P}_{tj} X_{t\cdot}^2 \boldsymbol{\beta}_j = \sum_{t=1}^{I}\sum_{j=1}^{I} \mathbf{P}_{tj}(X_{t\cdot}\boldsymbol{\alpha}_j - X_{t\cdot}^2\boldsymbol{\beta}_j)$$

$$(\mathbf{P}\boldsymbol{\gamma})^\intercal \mathbf{B}\mathbf{B}^\intercal(\mathbf{P}\boldsymbol{\gamma}) = \sum_{i=1}^{I}\sum_{j=1}^{I} \boldsymbol{\gamma}_i(\mathbf{P}^\intercal\mathbf{B}\mathbf{B}^\intercal\mathbf{P})_{ij}\boldsymbol{\gamma}_j = \sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{s=1}^{I}\sum_{t=1}^{I} \boldsymbol{\gamma}_i\boldsymbol{\gamma}_j(\mathbf{B}\mathbf{B}^\intercal)_{st}\mathbf{P}_{si}\mathbf{P}_{tj}$$

$$\begin{aligned}(\mathbf{P}\boldsymbol{\alpha})^\intercal(T\mathbf{X}\mathbf{X}^\intercal - \mathbf{B}\mathbf{B}^\intercal)(\mathbf{P}\boldsymbol{\alpha}) &= \sum_{i=1}^{I}\sum_{j=1}^{I} \boldsymbol{\alpha}_i(\mathbf{P}^\intercal(T\mathbf{X}\mathbf{X}^\intercal - \mathbf{B}\mathbf{B}^\intercal)\mathbf{P})_{ij}\boldsymbol{\alpha}_j \\ &= \sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{s=1}^{I}\sum_{t=1}^{I} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j(T\mathbf{X}\mathbf{X}^\intercal - \mathbf{B}\mathbf{B}^\intercal)_{st}\mathbf{P}_{si}\mathbf{P}_{tj}\end{aligned}$$

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

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

and

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

This results in a MIQP problem taking the form:

$$\text{minimize/maximize: } \mathbf{R}^T\mathbf{M}\mathbf{R} + \mathbf{D}^T\mathbf{R}$$

$$\text{subject to: } \sum_{i=1}^{I} \mathbf{R}_{(s-1)I+i} = 1 \quad \forall s = 1, \cdots, I,$$

$$\sum_{s=1}^{I} \mathbf{R}_{(s-1)I+i} = 1 \quad \forall i = 1, \cdots, I$$

$$\mathbf{R}_{(s-1)I+i} \in \{0,1\}$$

where $\mathbf{R}$ is a vector of length $I^2$ decision variables.

More generally, to implement the algorithm for an imbalanced SW-CRT design, where an unequal number of clusters are randomized at each step, we formally define the allocation matrix

$$\Delta\mathbf{X} = \begin{bmatrix} \mathbf{X}_2 - \mathbf{X}_1 & \mathbf{X}_3 - \mathbf{X}_2 & \cdots & \mathbf{X}_T - \mathbf{X}_{T-1} \end{bmatrix}$$

and the grouping function $G : \mathbb{R}^I \to S_2 \times \cdots \times S_T$ as

$$G(\mathbf{v}) = (\mathbf{v}[\Delta\mathbf{X}_2], \cdots, \mathbf{v}[\Delta\mathbf{X}_T])$$

where $\mathbf{v}[\Delta\mathbf{X}_i]$ is the unordered set of elements within $\mathbf{v}$ that correspond to unit entries within $\Delta\mathbf{X}_i$. For example, consider an imbalanced SW-CRT design with $I = 6$ clusters and $T = 5$ time-points, where one cluster is randomized to treatment initiation at step 1, two clusters are randomized at step 2, one cluster is randomized at step 3, and two clusters are randomized at step 4. The treatment status matrix is

$$\mathbf{X} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and the corresponding allocation matrix is

$$\Delta\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Consider $\mathbf{n} = (10, 15, 20, 40, 45, 50)$ with maximizing solution

$$\mathbf{R} = (0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,1,0)$$

So that, an optimal order of clusters is $\mathbf{v}^* = (50, 20, 10, 40, 15, 45)$. Applying the grouping function, we see

$$G(\mathbf{v}^*) = (\{50\}, \{20, 10\}, \{40\}, \{15, 45\})$$

where the order within $\{\}$ does not matter since those clusters are allocated to initiate treatment at the same time-point. Note that there are a total of four permutations of this solution, which will attain the same maximum power.

To improve performance of the optimization software, extra constraints can be added. In general, suppose $q_2, q_3, \cdots, q_T$ clusters, with $q_2 + q_3 + \cdots + q_T = I$, are randomized to initiate the intervention at time-points $2, 3, \cdots, T$, respectively. Then the order of the randomization at the same time-point is irrelevant. Additional constraints can be specified as follows

$$\text{subject to: } \sum_{i=1}^{I} i(\mathbf{R}_{(s-1)I+i} - \mathbf{R}_{sI+i}) \leqslant 0 \quad \forall s \in \{I, \cdots, I\} \setminus \{q_2, q_2 + q_3, \cdots, q_2 + \cdots + q_T\}$$

These constraints effectively tell the optimization software to find the solution in increasing index order $i$ for $n_i$ for time-points where more than one cluster is allocated to initiate the intervention. For the above example with $\mathbf{n} = (n_1, n_2, n_3, n_4, n_5, n_6) = (10, 15, 20, 40, 45, 50)$, the maximizing solution with these additional constraints will be

$$\mathbf{R} = (0,0,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,1,0)$$

So that, the optimal order of clusters found by the optimization software is $\mathbf{v}^* = (50, 10, 20, 40, 15, 45)$ since we observe

$$G(\mathbf{v}^*) = (\{n_6\}, \{n_1, n_3\}, \{n_4\}, \{n_2, n_5\}) = (\{50\}, \{10, 20\}, \{40\}, \{15, 45\})$$

has increasing indices in $n_i$ within each $\{\}$.

## Web Appendix D

Let's consider a balanced SW-CRT design with treatment status matrix $\mathbf{X}$, $q$ clusters randomized at each step, $b$ 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 $j = 1, \cdots, T$,

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

where $\lceil \cdot \rceil$ is the ceiling function. Now let's derive the joint distribution of $(X_{ij}, X_{lm})$. Assume without loss of generality that $j < m$. Then,

$$\mathbb{P}(X_{ij} = X_{lm} = 1) = \mathbb{P}(X_{ij} = 1)\mathbb{P}(X_{lm} = 1 | X_{ij} = 1) = \frac{\lceil \frac{j-b}{t} \rceil q}{I} \frac{\lceil \frac{m-b}{t} \rceil q - 1}{I - 1}$$

If $j > m$, the variables would change places, hence in general,

$$\mathbb{P}(X_{ij} = X_{lm} = 1) = \frac{(\lceil \frac{j-b}{t} \rceil \wedge \lceil \frac{m-b}{t} \rceil)q}{I} \frac{(\lceil \frac{j-b}{t} \rceil \vee \lceil \frac{m-b}{t} \rceil)q - 1}{I - 1}$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

When there is one baseline time-point and one time-point after each step (i.e. $b = t = 1$), these equations simplify,

$$\mathbb{P}(X_{ij} = 1) = \frac{(j-1)q}{I}, \ \mathbb{P}(X_{ij} = X_{lm} = 1) = \frac{\{(j-1) \wedge (m-1)\}q}{I} \frac{\{(j-1) \vee (m-1)\}q - 1}{I - 1}$$

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

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

so,

$$\mathbb{E}(\ell-z) = \frac{T-b+t}{2}\left(f+\frac{g}{3}(T+2b-t)\right)$$

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

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

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,

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

$$= \sum_{p=-1}^{I/q-1} \frac{(p+1)q}{I} \frac{(p+1)q-1}{I-1} \left\{ \frac{t(t-1)}{2} \right\} + \sum_{p=-1}^{I/q-1} \sum_{p'=p+1}^{I/q-1} \frac{(p+1)q}{I} \frac{(p'+1)q-1}{I-1} t^2$$
$$= \frac{(t-1)(T-b+t)(2I+q-3)}{12(I-1)} + \frac{(T-b+t)(T-b-t)(3I+2q-4)}{24(I-1)}$$

and

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

so,

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

Finally,

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

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 $N_{SW} = \sum_{i=1}^{I} n_i$ denote the number of participants sampled at each time-point and let $\kappa$ denote the sample coefficient of variation (CV) for cluster sizes $n_i$, so 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

$$\mathbb{E}(Tw - \ell^2) = \frac{T-b+t}{\sigma_e^4} \left[ \frac{N_{SW}^2}{I} \left( \frac{I-1}{I}\kappa^2 + 1 \right) Y_1 + \left\{ N_{SW}^2 - \frac{N_{SW}^2}{I} \left( \frac{I-1}{I}\kappa^2 + 1 \right) \right\} Y_2 \right]$$
$$= \frac{(T-b+t)N_{SW}^2}{\sigma_e^4} \left\{ \frac{(I-1)(Y_1 - Y_2)}{I^2}\kappa^2 + Y_2 + \frac{Y_1 - Y_2}{I} \right\}$$

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

These expectations simplify when $b = t = 1$,

$$\mathbb{E}(\ell - z) = \frac{T}{2}\left\{ f + \frac{g}{3}(T+1) \right\}, \qquad \mathbb{E}(y^2) = \frac{T}{12(I-1)}\left[ s_1 I(T-2) + f^2\{3IT - 2(2T-1)\} \right],$$
$$\mathbb{E}(Tw - \ell^2) = \frac{T(T+1)(f+gT)^2}{12(T-1)} \left\{ \frac{(T-2)}{I}\kappa^2 + T \right\}$$

## Web Appendix E

In order to obtain a 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:

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

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,

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

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.

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

where

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

These expressions simplify when $b = t = 1$,

$$C = \frac{T(T+1)}{12(T-1)} \left\{ \frac{(T-2)}{I}\kappa^2 + T \right\}, \quad U = \frac{IT}{2}, \quad V = \frac{IT(2T-1)}{6}$$

## 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,

$$\begin{aligned}\mathbb{E}_P\{\text{Var}(\widehat{\theta}|P)\} &\approx \frac{IT\sigma^2(\sigma^2+T\tau^2)}{\sigma^2\Big[IT\frac{I(T-b+t)}{2} - \Big\{\frac{I(T-b+t)}{2}\Big\}^2 - I^2C\Big] + T\tau^2\Big\{IT\frac{I(T-b+t)}{2} - I\frac{I(T-b+t)(2T-2b+t)}{6} - I^2C\Big\}} \\ &= \frac{12(T-b)T\sigma^2(\sigma^2+T\tau^2)}{I(T-b+t)(T-b-t)\{\sigma^2(2T-\frac{T+b}{I}\kappa^2)+T\tau^2(T+b-\frac{T+b}{I}\kappa^2)\}}\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_P\{\text{Var}(\widehat{\theta}|P)\} &\approx \frac{12T(T-b)\frac{\sigma_t^2(1-\rho)}{n}\big\{\frac{\sigma_t^2(1-\rho)}{n}+T\rho\sigma_t^2\big\}}{I(T-b+t)(T-b-t)\big\{\frac{\sigma_t^2(1-\rho)}{n}(2T-\frac{T+b}{I}\kappa^2)+T\rho\sigma_t^2(T+b-\frac{T+b}{I}\kappa^2)\big\}} \\ &= \frac{12T(T-b)\sigma_t^2(1-\rho)\{1+(Tn-1)\rho\}}{In(T-b+t)(T-b-t)[T\{2(1-\rho)+(T+b)n\rho\}-\frac{T+b}{I}\kappa^2\{1+(Tn-1)\rho\}]}\end{aligned}$$

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

$$\frac{\mathbb{E}_P\{\text{Var}(\widehat{\theta}|P)\}}{4\sigma_t^2/(nIT)} \approx \frac{3T(T-b)(1-\rho)\{1+(Tn-1)\rho\}}{(T-b+t)(T-b-t)\left[2(1-\rho)+(T+b)n\rho-\frac{T+b}{IT}\kappa^2\{1+(Tn-1)\rho\}\right]}$$

This design effect simplifies when $b = t = 1$,

$$DE_{w,\kappa} = \frac{3(T-1)(1-\rho)\{1+(Tn-1)\rho\}}{(T-2)\left[2(1-\rho)+(T+1)n\rho-\frac{T+1}{IT}\kappa^2\{1+(Tn-1)\rho\}\right]}$$

## Web Appendix G

When $\kappa = 0$, the design effect is

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

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

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

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

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

## Web Appendix H

$$
\begin{aligned}
RE &\approx \frac{\frac{3T(T-b)(1-\rho)\{1+(Tn-1)\rho\}}{(T-b+t)(T-b-t)\{2(1-\rho)+(T+b)n\rho\}}}{\frac{3T(T-b)(1-\rho)\{1+(Tn-1)\rho\}}{(T-b+t)(T-b-t)\left[2(1-\rho)+(T+b)n\rho-\frac{T+b}{IT}\kappa^2\{1+(Tn-1)\rho\}\right]}} \\
&= \frac{2(1-\rho)+(T+b)n\rho-\frac{T+b}{IT}\kappa^2\{1+(Tn-1)\rho\}}{2(1-\rho)+(T+b)n\rho} \\
&= 1-\frac{(T+b)\kappa^2\{1+(Tn-1)\rho\}}{IT\{2(1-\rho)+(T+b)n\rho\}} \\
&= 1-\frac{\kappa^2}{I}\left(1-\frac{(T-b)(1-\rho)}{T[2+\{(T+b)n-2\}\rho]}\right)
\end{aligned}
$$

When $b=t=1$ this simplifies to

$$
RE \approx 1-\frac{\kappa^2}{I}\left(1-\frac{(T-1)(1-\rho)}{T[2+\{(T+1)n-2\}\rho]}\right)
$$

## Web Appendix I

The power is

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

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

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

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

When $b=t=1$ this simplifies to

$$
TN_{SW} = T(DE_w N_{ind}+CF)
$$

- $N_{ind} = 4(\sigma_e^2+\tau^2)(z_{1-\beta}+z_{1-\alpha/2})^2/\theta_A^2$ is the total sample size required for an individually randomized trial with an anticipated treatment effect of $\theta_A$
- $DE_w = [3(T-1)(1-\rho)\{1+(Tn-1)\rho\}]/(T(T-2)[2+\{(T+1)n-2\}\rho])$ 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 $AT = (T-1)(1-\rho)/(T[2 + \{(T+1)n - 2\}\rho])$

## Web Appendix J

An R function to calculate the variance using the formula in equation 1 as well as the resulting power using a Wald test is given below.

```
power_ord <- function(n,tau_sq,sigma_e_sq,I,T,X_ij,effect_size,alpha)
{
  N_sw <- sum(n)
  f_plus_gT <- N_sw/sigma_e_sq
  f <- sum(n/(sigma_e_sq+(n*T*tau_sq)))
  numerator <- f*T*f_plus_gT
  l <- sum(t(n) %*% X_ij)/sigma_e_sq
  z <- (tau_sq/sigma_e_sq)*sum(rowSums(X_ij)^2*n^2*(1/(sigma_e_sq+n*T*tau_sq)))
  l_z <- l-z
  y <- sum(rowSums(X_ij)*n*(1/(sigma_e_sq+n*T*tau_sq)))
  y_sq <- y^2
  w <- sum((t(n)%*%X_ij)^2)/sigma_e_sq^2
  Tw_lsq <- T*w - l^2
  denominator_ord <- numerator*l_z - f_plus_gT*y_sq - f*Tw_lsq
  var_ord <- numerator/denominator_ord
  power_ord <- pnorm(effect_size/sqrt(var_ord) - qnorm(1 - alpha/2))
  return(c(var_ord,power_ord))
}
X.ij <- matrix(c(0,rep(1,4),rep(0,2),rep(1,3),rep(0,3),rep(1,2),rep(0,4),1),byrow=TRUE,nrow=4,ncol=5)
power_ord(c(10,15,45,50),0.05,0.95,4,5,X.ij,0.4,0.05)
[1] 0.02338494 0.74401069
```

Example code utilizing Gurobi in R is given below.

```
library(slam)
library(gurobi)
I <- 4
T <- 5
X.ij <- matrix(c(0,rep(1,4),rep(0,2),rep(1,3),rep(0,3),rep(1,2),rep(0,4),1),byrow=TRUE,nrow=4,ncol=5)
tau_sq <- 0.05
sigma_e_sq <- 0.95
```

```
n <- c(10,15,45,50)
N_sw <- sum(n)
f <- sum(n/(sigma_e_sq+(n*T*tau_sq)))
f_plus_gT <- N_sw/sigma_e_sq
alpha <- n/sigma_e_sq
beta <- n^2*(tau_sq/(sigma_e_sq*(sigma_e_sq+n*tau_sq*T)))
gamma <- n/(sigma_e_sq+n*tau_sq*T)
X_term1 <- T*X.ij%*%t(X.ij) - rowSums(X.ij)%*%t(rowSums(X.ij))
X_term2 <- rowSums(X.ij)%*%t(rowSums(X.ij))
X_term3 <- rowSums(X.ij)
X_term4 <- rowSums(X.ij)^2
R_1 <- matrix(NA,nrow=I^2,ncol=I^2)
R_2 <- matrix(NA,nrow=I^2,ncol=I^2)
D_1 <- rep(NA,I^2)
D_2 <- rep(NA,I^2)
j <- -(I-1)
for (i in 1:I){
  j <- j+I
  R_1[j:(j+I-1),] <- cbind(X_term1[i,1]*alpha%*%t(alpha),X_term1[i,2]*alpha%*%t(alpha),
                           X_term1[i,3]*alpha%*%t(alpha),X_term1[i,4]*alpha%*%t(alpha))
  R_2[j:(j+I-1),] <- cbind(X_term2[i,1]*gamma%*%t(gamma),X_term2[i,2]*gamma%*%t(gamma),
                           X_term2[i,3]*gamma%*%t(gamma),X_term2[i,4]*gamma%*%t(gamma))
  D_1[j:(j+I-1)]   <- X_term3[i]*alpha
  D_2[j:(j+I-1)]   <- X_term4[i]*beta
}
M <- -(f_plus_gT)*R_2 -f*R_1
D <- f*T*f_plus_gT*(D_1-D_2)
model <- list()
model$A          <- matrix(c(rep(1,I),rep(0,(I-1)*I),
                             rep(0,I),rep(1,I),rep(0,(I-2)*I),
                             rep(0,2*I),rep(1,I),rep(0,I),
                             rep(0,3*I),rep(1,I),
                             rep(c(1,rep(0,I-1)),4),
                             rep(c(0,1,rep(0,I-2)),4),
                             rep(c(0,0,1,rep(0,I-3)),4),
                             rep(c(0,0,0,1),4)
                             ), nrow=2*I, ncol=I^2, byrow=T)
```

```
model$Q          <- M
model$obj        <- D
model$modelsense <- 'min'
model$rhs        <- rep(1,2*I)
model$sense      <- rep('=',2*I)
model$vtype      <- 'B'
params <- list()
params$OutputFlag <- 0
# Sequence that will give minimum power
min_power <- gurobi(model,params)
which_order <- which(min_power$x %in% 1) %% I
which_order[which_order==0] <- I
n[which_order]
[1] 10 45 50 15
# Sequence that will give maximum power
model$modelsense <- 'max'
max_power <- gurobi(model,params)
which_order <- which(max_power$x %in% 1) %% I
which_order[which_order==0] <- I
n[which_order]
[1] 45 15 10 50
```

An R function to calculate the variance using the formula in equation 2 as well as the resulting power using a Wald test is given below. For equation 2, set $b = t = 1$. For an extension when $b \neq 1$ or $t \neq 1$, see Web Appendix D.

```
power_q <- function(n,tau_sq,sigma_e_sq,I,q,b,t,effect_size,alpha)
{
  T <- (I/q)*t + b
  N_sw <- sum(n)
  f_plus_gT <- N_sw/sigma_e_sq
  f <- sum(n/(sigma_e_sq+(n*T*tau_sq)))
  numerator <- f*T*f_plus_gT
  g <- sum((tau_sq*n^2)/(sigma_e_sq*(sigma_e_sq+T*tau_sq*n)))
  E_l_z <- ((T-b+t)/2)*(f+g*((T+2*b-t)/3))
  s1 <- sum((n/(sigma_e_sq+n*T*tau_sq))^2)
  E_y_sq <- ((T-b+t)/(12*(I-1)))*(I*(T-b-t)*s1+f^2*(3*I*(T-b+t)-2*(2*T-2*b+t)))
  CV_sq <- var(n)/(mean(n)^2)
  CV_term <- (T+b)*(T-b-t)/I
```

```
  E_Tw_lsq <- ((T-b+t)*f_plus_gT^2/(12*(T-b)))*(CV_sq*CV_term+T^2+2*b*T-t*T-3*b^2+3*b*t)
  var_q <- numerator/((numerator*E_l_z) - (f_plus_gT*E_y_sq) - (f*E_Tw_lsq))
  power_q <- pnorm(effect_size/sqrt(var_q) - qnorm(1 - alpha/2))
  return(c(var_q,power_q))
}
power_q(c(10,15,45,50),0.05,0.95,4,1,1,1,0.4,0.05)
[1] 0.02210032 0.76752113
```

An R function to calculate the variance using the formula in equation 3 as well as the resulting power using a Wald test is given below. For equation 3, set $b = t = 1$. For an extension when $b \neq 1$ or $t \neq 1$, see Web Appendix E.

```
power_app_q <- function(n_avg,CV,tau_sq,sigma_e_sq,I,q,b,t,effect_size,alpha)
{
  T <- (I/q)*t + b
  f_plus_gT <- (I*n_avg)/sigma_e_sq
  numerator_CV <- I*T*(sigma_e_sq/n_avg)*(sigma_e_sq/n_avg + (tau_sq*T))
  U <- (I*(T-b+t))/2
  V <- (I*(T-b+t)*(2*T-2*b+t))/6
  CV_term <- (T+b)*(T-b-t)/I
  C <- ((T-b+t)/(12*(T-b)))*((CV^2)*CV_term+T^2+2*b*T-t*T-3*b^2+3*b*t)
  denominator_CV <- (sigma_e_sq/n_avg)*(I*T*U-U^2-I^2*C) + tau_sq*T*(I*T*U-I*V-I^2*C)
  var_app_q <- numerator_CV/denominator_CV
  power_app_q <- pnorm(effect_size/sqrt(var_app_q) - qnorm(1 - alpha/2))
  return(c(var_app_q,power_app_q))
}
power_app_q(30,0.6804138,0.05,0.95,4,1,1,1,0.4,0.05)
[1] 0.02200885 0.76922380
```

An R function to calculate the sample size using the formula in equation 6 is given below. For equation 6, set $b = t = 1$. For an extension when $b \neq 1$ or $t \neq 1$, see Web Appendix I.

```
N_sw <- function(power,effect_size,alpha,n_avg,CV,sigma_t_sq,rho,T,b,t)
{
  N_ind <- 4*sigma_t_sq*(qnorm(power) + qnorm(1-alpha/2))^2/effect_size^2
  DE_w <- (3*(T-b)*(1-rho)*(1+(T*n_avg-1)*rho))/((T-b+t)*(T-b-t)*(2*(1-rho)+(T+b)*n_avg*rho))
  CF <- n_avg*(CV^2)*(1-((T-b)*(1-rho))/(T*(2*(1-rho)+(T+b)*n_avg*rho)))
  N_sw <- DE_w*N_ind + CF
  T_N_sw <- T*N_sw
  I <- N_sw/n_avg
  clusters_per_step <- I/((T-b)/t)
```

```
  return(c(T_N_sw,N_sw,I,clusters_per_step))
}
N_sw(0.8,0.267,0.05,100,1.4,1,0.05,3,1,1)
[1] 2399.249599  799.749866    7.997499    3.998749
```

This software code is also available as a zip-file in a web supplement.

## Web Appendix K

For the generalized model described in Hooper et al. (2016) and Li et al. (2018), the variance matrix of the cluster-level mean responses is $\mathbf{V}^* \in \mathbb{R}^{IT\times IT}$, which is block-diagonal, $\mathbf{V}^* = \text{diag}(\mathbf{V}_1^*, \cdots, \mathbf{V}_I^*)$, with each block

$$\mathbf{V}_i^* = (\sigma_i^2 + \delta^2)\mathbf{I}_T + \tau^2 \mathbf{1}_T \mathbf{1}_T^\intercal$$

Therefore,

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

where

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

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

Noting the identity $f^* + g^*T = \sum_{i=1}^I \frac{1}{\sigma_i^2+\delta^2}$ and defining $w^* = \sum_{j=1}^T \left(\sum_{i=1}^I \frac{X_{ij}}{\sigma_i^2+\delta^2}\right)^2$, we have the following generalization of equation 1:

$$\text{Var}^*(\widehat{\theta}|P=p) = [(\mathbf{Z}^\intercal(\mathbf{V}^*)^{-1}\mathbf{Z})^{-1}]_{(T+1),(T+1)} = \frac{f^*T(f^*+g^*T)}{f^*T(f^*+g^*T)(\ell^*-z^*) - (f^*+g^*T)(y^*)^2 - f^*\{Tw^* - (\ell^*)^2\}}$$

To find the upper and lower bounds for the power across all possible randomizations, the maximum and minimum of the generalized denominator $D^*(\mathbf{v}) = f^*T(f^*+g^*T)\left(\ell^*-z^*\right) - \left(f^*+g^*T\right)\left(y^*\right)^2 - f^*\left\{Tw^* - (\ell^*)^2\right\}$ is sought. This results in the following MIQP problem:

$$\begin{aligned} \text{minimize/maximize: } & \mathbf{R}^T\mathbf{M}^*\mathbf{R} + (\mathbf{D}^*)^T\mathbf{R} \\ \text{subject to: } & \sum_{i=1}^I \mathbf{R}_{(s-1)I+i} = 1 \quad \forall s = 1,\cdots,I \\ & \sum_{s=1}^I \mathbf{R}_{(s-1)I+i} = 1 \quad \forall i = 1,\cdots,I \\ & \mathbf{R}_{(s-1)I+i} \in \{0,1\} \end{aligned}$$

where

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

$$\mathbf{M}^*_{(s-1)I+i,(t-1)I+j} = -(f^* + g^*T)\left(\sum_{k=1}^{T} X_{sk}\right)\left(\sum_{k=1}^{T} X_{tk}\right)\frac{1}{(\sigma_i^2+\tau^2T+\delta^2)(\sigma_j^2+\tau^2T+\delta^2)} - f^*\left\{T\sum_{k=1}^{T} X_{sk}X_{tk} - \left(\sum_{k=1}^{T} X_{sk}\right)\left(\sum_{k=1}^{T} X_{tk}\right)\right\}\frac{1}{(\sigma_i^2+\delta^2)(\sigma_j^2+\delta^2)}$$

  for $s, t, i, j = 1, \cdots, I$
- $\mathbf{D}^*$ is a vector of length $I^2$ with elements

$$\mathbf{D}^*_{(t-1)I+j} = f^*T(f^* + g^*T)\left\{\left(\sum_{k=1}^{T} X_{tk}\right)\frac{1}{\sigma_j^2+\delta^2} - \left(\sum_{k=1}^{T} X_{tk}\right)^2 \frac{\tau^2}{(\sigma_j^2+\delta^2)(\sigma_j^2+\tau^2T+\delta^2)}\right\}$$

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

We calculate $\mathbb{E}\left(\ell^* - z^*\right)$, $\mathbb{E}\left\{(y^*)^2\right\}$ and $\mathbb{E}\left\{Tw^* - (\ell^*)^2\right\}$ as:

$$\mathbb{E}(\ell^* - z^*) = \frac{T-b+t}{2}\left\{f^* + \frac{g^*}{3}(T+2b-t)\right\}$$

Defining $s_1^* = \sum_{i=1}^{I} \frac{1}{(\sigma_i^2+T\tau^2+\delta^2)^2}$.

$$\mathbb{E}\left\{(y^*)^2\right\} = \frac{T-b+t}{12(I-1)}\left[(T-b-t)Is_1^* + \{3I(T-b+t) - 2(2T-2b+t)\}(f^*)^2\right]$$

$$\mathbb{E}\left\{Tw^* - (\ell^*)^2\right\} = (T-b+t)\left[Y_1\left\{\sum_{i=1}^{I}\frac{1}{(\sigma_i^2+\delta^2)^2}\right\} + Y_2\left\{2\sum_{i<i'}\frac{1}{(\sigma_i^2+\delta^2)(\sigma_{i'}^2+\delta^2)}\right\}\right]$$

where

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

These expectations simplify when $b = t = 1$,

$$\mathbb{E}(\ell^* - z^*) = \frac{T}{2}\left\{f^* + \frac{g^*}{3}(T+1)\right\}$$

$$\mathbb{E}\left\{(y^*)^2\right\} = \frac{T}{12(I-1)}\left[(T-2)Is_1^* + \{3IT - 2(2T-1)\}(f^*)^2\right]$$

$$\mathbb{E}\left\{Tw^* - (\ell^*)^2\right\} = T\left[Y_1\left\{\sum_{i=1}^{I}\frac{1}{(\sigma_i^2+\delta^2)^2}\right\} + Y_2\left\{2\sum_{i<i'}\frac{1}{(\sigma_i^2+\delta^2)(\sigma_{i'}^2+\delta^2)}\right\}\right]$$

where

$$Y_1 = \frac{T+1}{6} \qquad \text{and} \qquad Y_2 = \frac{IT+2qT-2T-2}{12(I-1)}$$

## Web Appendix L

## Simulation Study Design

We conducted three main simulations studies to evaluate the performance of the proposed

methods. Firstly, the method to determine the power for each randomization order was evaluated. Secondly, the expected power across all randomization orders was evaluated in detail under four study design scenarios with varying degrees of cluster size variation. Thirdly, a method that simply plugs-in the harmonic mean cluster size was assessed.

To firstly evaluate the impact of the order of randomization for particular known cluster sizes, we considered a design with 6 clusters ($I = 6$) with one cluster randomized to the intervention at each step ($q = 1$), and a continuous outcome. The cluster sizes were 4, 11, 18, 21, 22 and 104, resulting in an arithmetic mean cluster size of 30 and a CV of $\kappa = 1.23$. The effect size ($\theta_A$) was set so that the power calculated using a fixed cluster size of 30 by equation 8 of Hussey and Hughes (2007) was 80%. Then, the power was calculated for each of the $6! = 720$ possible randomization sequences of the varying cluster sizes using the variance formula in equation 1. The method described in Section 2.3 was used to determine upper and lower bounds for the power across all randomization sequences. Furthermore, for this design, the expected power accounting for cluster size variation when all cluster sizes are known was calculated using the variance formula in equation 2. Additionally, the approximate expected power when only a cluster size arithmetic mean and CV ($\kappa$) is known prior to randomization using equation 3 was calculated. To simulate the empirical power, for each randomization sequence the data were generated using the model given in Section 2.1. For convenience, both $\mu$ and $\beta_j$ for $j = 1, .., T-1$ were set at zero. 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. The intra-cluster correlation, $\rho$, was set at 0.05. Data were analyzed by the same linear mixed effect model using the "lmer" function from the "lme4" R package. A two-tailed Wald test for the treatment effect was generated, and the empirical power calculated by the proportion of simulated results from 3,500 replications for each randomization sequence with a p-value $< 0.05$. This evaluation method was repeated

for two more scenarios; firstly using the same design and model in Section 2.1, but with cluster sizes of 4, 11, 18, 21, 62 and 64, resulting in an arithmetic mean cluster size of 30 and a CV of $\kappa = 0.87$, and secondly using the same design with cluster sizes of 4, 11, 18, 21, 22 and 104, but using the generalized model in Section 2.7 with $\sigma_e^2 = 0.95$, $\tau^2 = 0.04$ and $\delta^2 = 0.01$.

To secondly conduct a detailed evaluation of the variance formulas in equation 2 and 3 to estimate the expected power under cluster size variation, we considered four different design scenarios:

(1) A continuous outcome with 4 clusters ($I = 4$) with 1 cluster randomized at each step ($q = 1$).
(2) A continuous outcome with 6 clusters ($I = 6$) with 1 cluster randomized at each step ($q = 1$).
(3) A continuous outcome with 12 clusters ($I = 12$) with 3 clusters randomized at each step ($q = 3$), with $b = 2$ baseline time-points and $t = 3$ time-points between each step.
(4) A count outcome with 6 clusters ($I = 6$) with 1 cluster randomized at each step ($q = 1$).

For each simulation scenario, the total number of participants contributing data ($N_{SW}$) at each time-point was kept fixed, so that the arithmetic mean cluster size was 30. For example, for the design with 4 clusters, there were $N_{SW} = 120$ participants contributing data at each time-point. Under cluster size variation, the number of participants contributing data from each cluster, i.e. $n_i$ for $i = 1, ..., I$ at every time-point 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 ($\kappa$) ranged from 0 to a maximum of 3.5. The effect size ($\theta_A$) was set so that the power calculated based on a fixed cluster size of 30 would be 80%. Then the expected power accounting for cluster size variation assuming all cluster sizes are known and with only knowledge of the cluster size arithmetic mean and CV ($\kappa$) were calculated using the variance formulas in equation 2 and 3, respectively. To simulate the empirical power, the variable size clusters were placed in a random order with an equal number of clusters (denoted by $q$) randomized to initiate the intervention at each step, and then data were simulated using the model given in Section 2.1. All 3,500 simulation replications used the same random order. Both $\mu$ and $\beta_j$ for $j = 1, .., T-1$ were set at zero. For continuous outcomes, the total variance $\sigma_t^2 = \sigma_e^2 + \tau^2$ was fixed at 1 and $\rho$ at 0.05. Data were analyzed by the same linear mixed effect model. For simulations involving count outcomes, $\sigma^2 = (1 + e^{\theta_A})/2$ and $\tau^2 = (\rho\sigma^2)/(1-\rho)$. $\alpha_i$ 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". A two-tailed Wald test for the treatment effect was generated, and the empirical power calculated by the proportion of simulated results from 3,500 replications with p-value $< 0.05$. 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. Using 3,500 simulations resulted in a Monte Carlo error of $\leqslant 0.75\%$. Simulations were repeated using the generalized model in Section 2.7 for a design with a continuous outcome and 6 clusters ($I = 6$) with one randomized at each step ($q = 1$).

To lastly address the question of whether the power calculation can simply be based on the harmonic mean with no further adjustment we produced some representative examples of cluster sizes that held the harmonic mean fixed at 30, but varied the arithmetic mean and CV. We used the design with a continuous outcome, 6 clusters ($I = 6$) with one cluster randomized at each step ($q = 1$), and an effect size ($\theta_A$) so that plugging-in the harmonic mean of 30 gave 80% power.

**Additional Simulation Study Results**

Supplementary figure S1 evaluated clusters of size 4, 11, 18, 21, 62 and 64, under the model in Section 2.1.

[Figure 1 about here.]

Supplementary figures S2 and S3 confirm the accuracy of the formulas for the generalization specified in Section 2.7. Supplementary figure S2 is for clusters of size 4, 11, 18, 21, 22 and 104.

[Figure 2 about here.]

[Figure 3 about here.]

**Performance of the Algorithm to Optimize the Treatment Effect Variance Over All Possible Randomization Sequences**

For a design with 8 ($I = 8$) or 10 ($I = 10$) clusters where one cluster is randomized at each step ($q = 1$), there would be $40,320$ or $3,628,800$ possible orders of randomization, respectively. On a personal computer for a representative example with an arithmetic mean cluster sizes of 30 and a cluster size CV of around 0.5, to find the order that would give the maximum

power took 0.34 and 1.65 seconds for $I = 8$ and $I = 10$, respectively. To find the order for that would give the minimum power took 2.60 and 283.50 seconds for $I = 8$ and $I = 10$, respectively. For the scenario with 12 clusters with 3 randomized at each step ($I = 12,\ q = 3$), there would be $19,958,400$ possible randomizations. For this example, we found that providing Gurobi with extra constraints, which are given in Web Appendix C, improved performance. These constraints inform the software package that the order of randomization within a step is irrelevant. For example, if three clusters of sizes 18, 20 and 25 are randomized to initiate the intervention at the first step, the constraint informs the software that it does not matter if their order is $(\{18, 20, 25\}, \{18, 25, 20\}, \{20, 18, 25\}, \{20, 25, 18\}, \{25, 18, 20\}, \{25, 20, 18\})$, since all combinations will result in the same power. With these constraints it took 0.01 seconds to identify the order that resulted in maximum power and 44.91 seconds to identify the order that resulted in minimum power. Even without the addition of these constraints Gurobi was able to identify the order that resulted in maximum power in 0.52 seconds, and find a heuristic solution for the minimum power within 5 seconds (which for this example was the optimum once the extra constraints were added).

## References

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.

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

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

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.

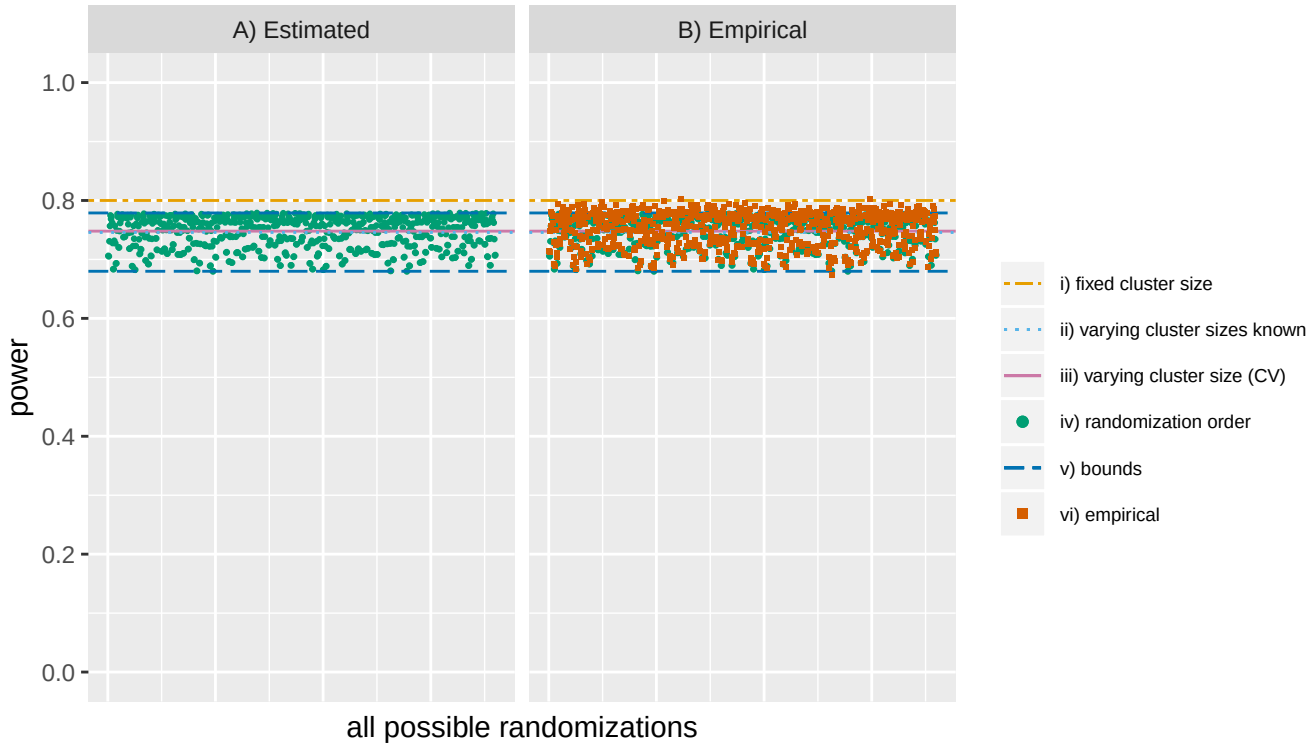


**Figure S1.** Power for all possible randomizations with upper and lower bounds indicated for a SW-CRT with 6 clusters of mean size ($n$) 30 and CV ($\kappa$) of 0.87. i) fixed cluster size: uses the variance formula in equation 8 of Hussey and Hughes (2007), ii) varying cluster sizes known: uses the variance formula in equation 2, iii) varying 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.

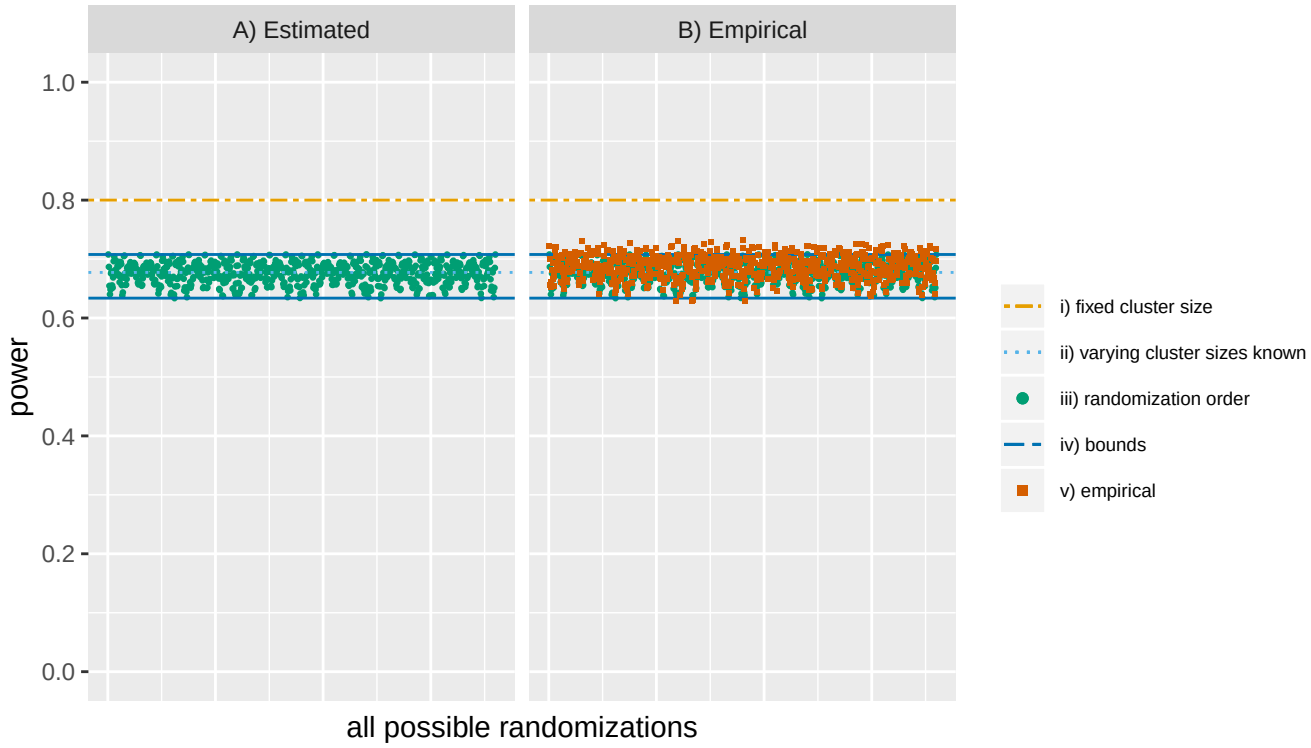


**Figure S2.** Power for all possible randomizations with upper and lower bound indicated for a SW-CRT with 6 clusters of mean size ($n$) 30, CV ($\kappa$) of 1.23 and $\sigma_e^2 = 0.95$, $\tau^2 = 0.04, \delta^2 = 0.01$. i) fixed cluster size: uses the variance formula in Li et al. (2018), ii) varying cluster sizes known, iii) randomization order and iv) bounds: use the formulas in Web Appendix K, v) empirical: the empirically simulated power from 3,500 simulations.

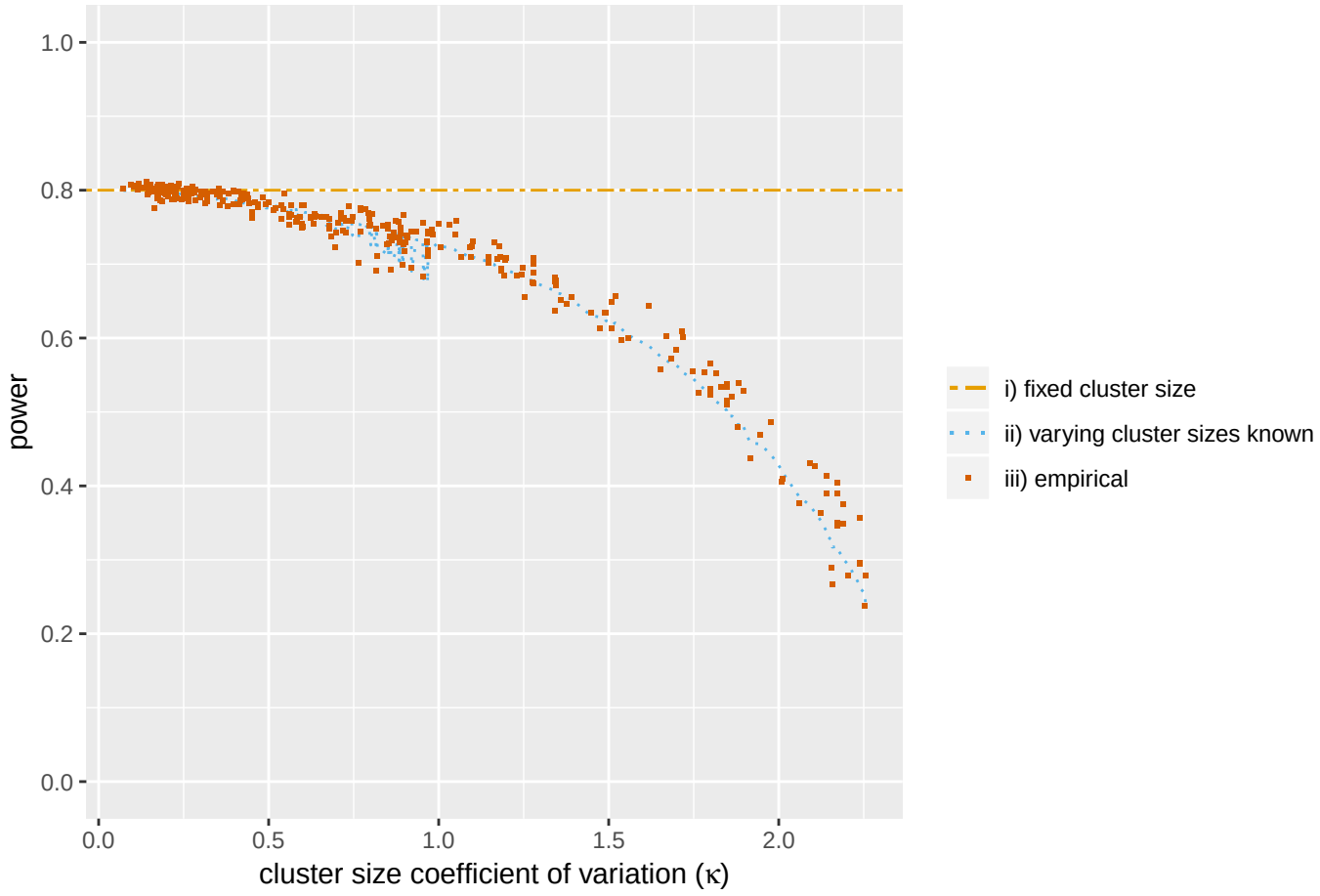


**Figure S3.** Estimated power as the cluster size coefficient of variation ($\kappa$) increases for a SW-CRT with 6 clusters of mean size (n) 30 and $\sigma_e^2 = 0.95$, $\tau^2 = 0.04, \delta^2 = 0.01$. i) fixed cluster size: uses the variance formula in Li et al. (2018), ii) varying cluster sizes known: uses the variance formula in Web Appendix K, iii) empirical: the empirically simulated power from 3,500 simulations.