Published in final edited form as:

*Stat Methods Med Res*. 2018 February ; 27(2): 480–489. doi:10.1177/0962280216632564.

# Sample size determinations for stepped-wedge clinical trials from a three-level data hierarchy perspective

**Moonseong Heo**[1], **Namhee Kim**[2], **Michael L Rinke**[3], and **Judith Wylie-Rosett**[1,4]

[1]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

[2]Department of Radiology, Albert Einstein College of Medicine, Bronx, NY, USA

[3]Department of Pediatrics, Children's Hospital at Montefiore, Albert Einstein College of Medicine, Bronx, NY, USA

[4]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA

## Abstract

Stepped-wedge (SW) designs have been steadily implemented in a variety of trials. A SW design typically assumes a three-level hierarchical data structure where participants are nested within times or periods which are in turn nested within clusters. Therefore, statistical models for analysis of SW trial data need to consider two correlations, the first and second level correlations. Existing power functions and sample size determination formulas had been derived based on statistical models for two-level data structures. Consequently, the second-level correlation has not been incorporated in conventional power analyses. In this paper, we derived a closed-form explicit power function based on a statistical model for three-level continuous outcome data. The power function is based on a pooled overall estimate of stratified cluster-specific estimates of an intervention effect. The sampling distribution of the pooled estimate is derived by applying a fixed-effect meta-analytic approach. Simulation studies verified that the derived power function is unbiased and can be applicable to varying number of participants per period per cluster. In addition, when data structures are assumed to have two levels, we compare three types of power functions by conducting additional simulation studies under a two-level statistical model. In this case, the power function based on a sampling distribution of a marginal, as opposed to pooled, estimate of the intervention effect performed the best. Extensions of power functions to binary outcomes are also suggested.

## Keywords

Stepped-wedge design; three level data; statistical power; sample size; design effect; effect size

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

**Corresponding author:** Moonseong Heo, Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Belfer 13th Floor, Bronx, NY 10461, USA. moonseong.heo@eisntein.yu.edu.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

# 1 Introduction

Stepped-wedge (SW) clinical trial design is a variation of cluster randomized trials and is a type of crossover design at the cluster level as treatment assignments are designed to be progressively crossed over unilaterally from a control to an experimental arm until all clusters are completely crossed over.[1] The random element of a SW design is the assignment of time points of the crossover to the clusters. The main advantage of the SW design is the relaxation of logistical constraints related to human or financial resources for conduct of classical cluster randomized trials,[1] although there are challenges in implementing in real-world settings.[2,3] The SW design is also useful when clinical equipoise is not met and it is unethical to randomize participants to the control arm for the length of the study. Further detailed discussion on this issue can be found in Prost et al.[4] Additional considerations that should be taken into account for conducting SW trials are suggested by Hargreaves et al.[5] The SW design has been steadily implemented in a variety of trials,[6] and a systematic review concerning characteristics of published SW trials is conducted by Brown and Lilford, 7 and more recently and comprehensively by Beard et al.[8]

As is the case for all types of randomized clinical trials, sample size determination is an indispensible element of the SW design. Hussey and Hughes[9] have proposed a widely used closed-form sample size determination formula for the SW design considering a random effects model. Woertman et al.[10] converted Hussey and Hughes' formula to a design effect of the SW design in comparison with a conventional two parallel arm design. Baio et al.[11] also suggested a design effect formula under a different setting and statistical model. Hemming et al.[12] evaluated the impact of intra-cluster correlations on statistical power or sample size through design effects under various types of SW designs. In addition, simulation studies for power analysis without explicit formula have also been conducted by Biao et al.[11] and Van den Heuvel et al.[13]

In all those derivations above, although the first level correlations (denoted below by $\rho_1$) of outcomes among participants in the same times or periods within the same clusters were taken into account, the second level correlations (denoted below by $\rho_2$) of outcomes among participants between times or periods within the same clusters were not explicitly considered for computing power or determining sample sizes. The latter correlations would need to be modeled in a statistical model for SW design trials because SW designs by definition naturally assume a three-level data hierarchy, as participants are nested within times or periods that are in turn nested within clusters in SW designs. The nomenclatures for units of levels should depend on the study context; for example, depending on research settings, the third level units can be physicians, clinics, hospitals, schools, communities, and districts to name a few. Hereafter, however, we refer to "cluster", "period", and "participant" as the third, the second, and the first level data units, respectively, in the SW design.

The primary aim of this paper is to derive explicit closed-form power functions which consider also the second level correlations by formulating a three-level model accounting for the two types of correlations. To this end, in section 2, we introduce a SW design with design parameter notations. In section 3, we specify the three-level model and formulate the two types of correlations. In section 4, we estimate an overall treatment effect by pooling cluster-specific effect estimates since the number of periods exposed to the experimental condition is not identical across all clusters. A power function is derived based on a sampling distribution of the pooled estimate of the overall treatment effect. In section 5, as a secondary aim, we compare performances of three power functions including that of Hussey and Hughes[9] under a two-level model when $\rho_2$ is assumed to be 0 as has previously been implicitly assumed. In section 6, simulation studies compare validities of all power functions under both two- and three-level models. Discussion follows in section 7.

## 2 Stepped wedge design parameters

Here we consider the SW design depicted in Figure 1, similar to that which was considered in Woertman et al.,[10] to illustrate design parameters. The total number of steps is represented by $S$ (≥1); the number of clusters in each step is represented by $c$ (≥1); the number of periods for each "depth" of a step under an experimental condition (gray areas) is represented by $p$ (≥1); and each cluster has $b$ (≥0) number of "baseline" periods under a control condition (blank areas). The clusters are indexed by $i = 1, 2, \ldots, I = cS$, this being the total number of clusters. Let us further denote by $S_{(i)}$ the depths of steps for the $i$th cluster under the experimental condition: e.g., $S_{(i)} = 1$ for $i = 1, 2$; $S_{(i)} = 2$ for $i = 3, 4$; and so on. The periods nested within clusters are indexed by $j = 1, 2, \ldots, J = b + pS$, this being the total number of periods per cluster. Study participants nested within each period are indexed by $k = 1, 2, \ldots, K$, this being referred to as "cell" size or the number of participants for each cell in Figure 1. Let us assume that the participants belong to only one cell without cross-over to other clusters or periods. Then the total number of participants $N$ will be $N = IJK = cS(b + pS)K$. The parameter values for the SW design depicted in Figure 1 can be found in its footnote. The sets of indices indicating observations from the experimental and control arms will be denoted by $E$ and $C$, respectively.

## 3 Statistical model for three level data structure

A statistical model for testing an experimental intervention/treatment effect under a SW design can be formulated as follows.

$$Y_{\mathrm{ijk}}=\beta_0+\delta X_{\mathrm{ijk}}+u_i+u_{j(i)}+e_{\mathrm{ijk}} \qquad (1)$$

The study outcome is denoted by $Y_{ijk}$ ($i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2\ldots, K$) and the experimental arm indicator is denoted by $X_{ijk} = 1$ for experimental arm, and=0 otherwise. Likewise, the control arm indicator is denoted by $W_{ijk}$=1 – $X_{ijk}$=1 for control arm, and =0 otherwise. Then, $X_{ijk}W_{i'j'k'} = 0$ if $i = i'$, $j = j'$, and $k = k'$; otherwise, the product is either 0 or 1 depending on the configurations of the indices. The two sets, E and C, are defined as $E = \{i, j, k \mid X_{ijk} = 1\} = \{i, j, k \mid W_{ijk} = 0\}$ and $C = \{i, j, k \mid W_{ijk} = 1\} = \{i, j, k \mid X_{ijk} = 0\}$.

The fixed-effect overall intercept is denoted by $\beta_0$, and the fixed experimental intervention effect by $\delta$ in model (1). The distribution of the cluster-level random intercepts $u_i$ is assumed to be normal as $u_i \sim N(0, \sigma_3^2)$ and so is that of the period-level random intercepts $u_{j(i)}$ as. $u_{j(i)} \sim N(0, \sigma_2^2)$. The distribution of the errors $e_{ijk}$ is also assumed to be normal as

$e_{\mathrm{ijk}} \sim N(\sigma_e^2)$. Among these random components, it is further assumed that $u_i \perp u_{j(i)} \perp e_{ijk}$, i.e., these three random components are mutually independent. In addition, *conditional independence* is assumed for all $u_{j(i)}$ and $e_{ijk}$, whereas the $u_i$ are *unconditionally* independent. That is, $u_{j(i)}$ are independent over $j$ conditional on $u_i$, and $e_{ijk}$ are independent over $k$ conditional on $u_i$, and $u_{j(i)}$.

Under model (1), and the elements of the covariance matrix are

$$\mathrm{Cov}(Y_{\mathrm{ijk}}, Y_{i'j'k'}) = 1(i=i' \& j=j' \& k=k')\sigma_e^2 + 1(i=i' \& j=j')\sigma_2^2 + 1(i=i')\sigma_3^2 \quad (2)$$

where 1(.) is an indicator function. It follows that $Var(Y_{ijk}) = Cov(Y_{ijk}, Y_{ijk}) = \sigma^2$, where $\sigma^2 = \sigma_e^2 + \sigma_2^2 + \sigma_3^2$. Therefore, the correlations among the level two data (i.e., among outcomes from different periods but the same cluster) can be expressed for $j \neq j'$ as follows:

$$\rho_2 = \sigma_3^2 / \sigma^2 \quad (3)$$

The correlations among the level one data (i.e., among outcomes measured from different participants in the same period within the same cluster) can be expressed for $k \neq k'$ as

$$\rho_1 = (\sigma_2^2 + \sigma_3^2) / \sigma^2 \quad (4)$$

As a result, $\rho_1$ is greater than or equal to $\rho_2$, that is, $\rho_1 \geq \rho_2$.

## 4 Estimate of intervention effect and power function

To estimate the overall intervention effect $\delta$, we consider each cluster as a stratum because outcome observations between periods within clusters are correlated and the numbers of periods exposed to control and experimental conditions are not necessarily identical across the clusters. This means that the variances of the cluster-specific effect estimates are not necessarily identical. In our approach, we first estimate an intervention effect for each cluster/stratum in a cluster-specific fashion, and then pool the cluster-specific estimates into an overall estimate of $\delta$ in model (1) by applying a fixed-effect meta-analytic approach[14] as $\delta$ is assumed to be fixed and homogenous across clusters.

The intervention effect for the $i$th cluster is denoted by $\delta_i$. A moment of estimate, $\tilde{\delta}_i$, of $\delta_i$ can then be obtained as $\tilde{\delta}_i = \tilde{\theta}_{i,E} - \tilde{\theta}_{i,C}$, where $\tilde{\theta}_{i,E} = \sum_j^J \sum_k^K X_{\mathrm{ijk}} Y_{\mathrm{ijk}} / N_{i,E}$ and $\tilde{\theta}_{i,C} = \sum_j^J \sum_k^K W_{\mathrm{ijk}} Y_{\mathrm{ijk}} / N_{i,C}$ are the means of outcome $Y$ for the participants in the experimental and control arms in the $i$th cluster, respectively; $N_{i,E}$ and $N_{i,C}$ represent the number of participants in the experimental and control arms in the $i$th cluster, respectively. That is, $N_{i,E} = \sum_j^J \sum_k^K X_{\mathrm{ijk}} = J_{i,E} K$ and $N_{i,C} = \sum_j^J \sum_k^K W_{\mathrm{ijk}} = J_{i,C} K$, where $J_{i,E} = \#\{j | X_{ajk} = 1, a = i\} = pS_{(i)}$ and $J_{i,C} = \#\{j / W_{ajk} = 1, a = i\} = b + p(S - S_{(i)})$ are the numbers of periods under the experimental and control conditions in the *i*th cluster; #{.} denotes the number of elements in the set {.}. It follows that $J_{i,E} + J_{i,C} = J$, $N_{i,E} = pS_{(i)}K$ and $N_{i,C} = bK + p(S - S_{(i)})K$. Under this setting, the variances of $\tilde{\theta}_{i,E}$ and $\tilde{\theta}_{i,C}$ and the covariance between them can be derived as follows:

$$\mathrm{Var}(\tilde{\theta}_{i,E})=\frac{\sigma^2}{J_{i,E}K}\{1+(K-1)\rho_1+K(J_{i,E}-1)\rho_2\},\ \mathrm{Var}(\tilde{\theta}_{i,C})=\frac{\sigma^2}{J_{i,C}K}\{1+(K-1)\rho_1+K(J_{i,C}-1)\rho_2\}$$

and $\mathrm{Cov}(\tilde{\theta}_{i,C},\tilde{\theta}_{i,E})=\sigma_3^2=\sigma^2\rho_2$. It follows that the variance of $\tilde{\delta}_i$ can be expressed as below:

$$\mathrm{Var}(\tilde{\delta}_i)=\mathrm{Var}(\tilde{\theta}_{i,E})+\mathrm{Var}(\tilde{\theta}_{i,C})-2\mathrm{Cov}(\tilde{\theta}_{i,E},\tilde{\theta}_{i,C})=\frac{fJ\sigma^2}{J_{i,E}J_{i,C}K}$$

where

$$f=1+(K-1)\rho_1-K\rho_2 \qquad (5)$$

which is the design effect for three-level trials that randomly assign treatments at the second level within clusters.[15,16] This design effect $f$ is an increasing function of $\rho_1$ and a decreasing function of $\rho_2$.

An estimate $\tilde{\delta}$ of overall intervention effect can now be obtained as a pooled estimate of $\tilde{\delta}_i$ ‘s weighted by their corresponding inverse variances as follows:

$$\tilde{\delta}=\sum_{i=1}^{I}\omega_i\tilde{\delta}_i/\sum_{i=1}^{I}\omega_i \qquad (6)$$

where $\omega_i=1/\mathrm{Var}(\tilde{\delta}_i)$. This pooled estimate is a weighted mean of $\tilde{\delta}_i$’s. It follows that

$$Var(\tilde{\delta})=1/\sum_{i=1}^{I}\omega_i=\frac{fJ\sigma^2}{K\sum_{i=1}^{I}J_{i,E}J_{i,C}} \qquad (7)$$

Under the setting depicted in Figure 1, the following equation is straightforward:

$$\sum_{i=1}^{I}J_{i,E}J_{i,C}=c\sum_{m=0}^{S-1}(pm+b)(S-m)p=\mathrm{cpS}(S+1)\left\{\frac{p(S-1)+3b}{6}\right\}$$

This equation enables a power function to be expressed in terms of design parameters as follows:

$$\varphi=\Phi\left(|\delta|\,/\,\sqrt{\mathrm{Var}(\tilde{\delta})}-z_{1-\alpha/2}\right)=\Phi\left(|\Delta|\,\sqrt{\frac{\mathrm{cpKS}(S+1)(pS-p+3b)}{6f(b+pS)}}-z_{1-\alpha/2}\right)\quad(8)$$

where Φ (.) is the cumulative density function of a standard normal distribution, $z_{1-\alpha/2}$ = $\Phi^{-1}(1-\alpha/2)$, $\Phi^{-1}(.)$ is the inverse of Φ(.), and $\Delta = \delta/\sigma$ which is known as standardized effect size or Cohen's *d*.[17] The statistical power increases with increasing Δ, *b*, *c*, *S* (or *I*), *p*, *K*, *α*, and $\rho_2$, all of which decrease the variance (7). However, the statistical power decreases with increasing $\rho_1$ which increases *f* and thus increases the variance (7).

## 5 Power under model for two-level data

If $\rho_2=\sigma_3^2/\sigma^2$(3) is assumed to be 0, then this assumption is equivalent to assuming $\sigma_3^2=0$, and reduces model (1) to a model

$$Y_{\mathrm{ijk}}=\beta_0+\delta X_{\mathrm{ijk}}+u_{j(i)}+e_{\mathrm{ijk}}\quad(9)$$

for a two-level data structure. Subsequently, $Cov(Y_{ijk}, Y_{i'j'k'})$ in equation (2) reduces to

$$\mathrm{Cov}(Y_{\mathrm{ijk}},\,Y_{i'j'k'})=1\,(i=i'\&j=j'\&k=k')\sigma_e^2+1\,(i=i'\&j=j')\sigma_2^2$$

and likewise $\sigma^2\equiv\sigma_e^2+\sigma_2^2$ and

$$\rho\equiv\rho_1=\sigma_2^2/\sigma^2\quad(10)$$

The statistical power expressed in equation (7) in Hussey and Hughes,[9] can be re-expressed, denoted here by $\varphi_{\mathrm{HH}}$, utilizing the equations in the supplements of Woertman et al.[10] as follows in terms of the design parameters depicted in Figure 1

$$\varphi_{\mathrm{HH}}=\Phi\left(|\Delta|\,\sqrt{\frac{\mathrm{cpKS}(S-1/S)\,\{1+\rho(\mathrm{pKS}/2+bK-1)\}}{6(1-\rho)\,\{1+\rho(\mathrm{pKS}+bK-1)\}}}-z_{1-\alpha/2}\right)\quad(11)$$

This function is not a monotone increasing or decreasing function of *ρ*. Furthermore, $\varphi_{\mathrm{HH}}$ cannot be defined if *ρ* = 1 although this is unrealistic to occur.

In addition, the statistical power *φ* (8) for the three-level model can straightforwardly be reduced to

$$\varphi_0=\Phi\left(|\Delta|\,\sqrt{\frac{\mathrm{cpKS}(S+1)(pS-p+3b)}{6f_0(b+pS)}}-z_{1-\alpha/2}\right)\quad(12)$$

where

$$f_0 = 1 + (K-1)\rho \qquad (13)$$

which is the same as $f$ (5) with $\rho_2$ (3) and $\rho_1$ (4) replaced by 0 and $\rho$ (10), respectively; $f_0$ (13) is the design effect for level-two data structure.[18] The statistical power $\varphi_0$ is in fact based on pooling of cluster-specific effect estimates weighted by the inverses of the cluster-specific variances of the estimates, and is a monotone decreasing function of $\rho$ as is the case for $\varphi$ (8) that decreases with $\rho_1$.

We note that the clusters, however, became nominal without any influence on statistical inference, since $\sigma_3^2 = 0$ is assumed. That is to say that the periods are no longer assumed to be nested within clusters although individual observations $Y$ are still assumed to be nested with periods. Therefore, statistical power $\varphi^{(2)}$ below can be based on a sampling distribution of a marginal estimate of $\delta$ in model (9) for two-level data as follows:

$$\tilde{\delta}_M = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{\text{ijk}} Y_{\text{ijk}} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{\text{ijk}} - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} W_{\text{ijk}} Y_{\text{ijk}} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} W_{\text{ijk}} \qquad (14)$$

It follows that the power $\varphi^{(2)}$ can be obtained as follows[19]

$$\varphi^{(2)} = \Phi\left(|\Delta| \sqrt{\frac{K}{f_0\left(1/N_E^{(2)} + 1/N_C^{(2)}\right)}} - z_{1-\alpha/2}\right) \qquad (15)$$

where $N_E^{(2)} = \#\left\{i, j \middle| X_{\text{ijk}} = 1\right\} = \text{pcS}(S+1)/2$ and $N_C^{(2)} = \#\left\{i, j \middle| W_{\text{ijk}} = 1\right\} = cS(b + p(S-1)/2)$ are the numbers of total periods for the experimental and control arms, respectively (the numbers of gray-colored and blank "cells" in Figure 1, respectively). It can be seen that the power function $\varphi^{(2)}$ is also a monotone decreasing function of $\rho$ (10).

## 6 Simulation study

We conducted simulations using the SAS v9.3 PROC MIXED routine with a restricted maximum likelihood fitting option to (1) validate the power function $\varphi$ (8) derived under the three-level model (1); and (2) compare three power functions $\varphi_{\text{HH}}$ (11), $\varphi_0$ (12), and $\varphi^{(2)}$ (15) under the two-level model (9). We note that it is possible to theoretically derive closed-form power functions with varying $K_{ij}$, the number of observations per period per cluster. However, it will be cumbersome not only to express exact formulas but also to compute power functions. Therefore, to assess applicability of the power functions under varying $K_{ij}$, we randomly drew $K_{ij}$ from uniform distributions $K_{ij} \sim U(a, b)$ with a = $K$ − floor(3 $K$/4) and b = $K$ + floor(3 $K$/4) so that $a > 0$ and $E\{U(a, b)\} = (a + b)/2 = K$, where floor($x$) returns the greatest integer smaller than or equal to $x$.

The magnitudes of all of theoretical power functions are compared with those of empirical power estimated from the simulations. To compute simulation-based empirical power, which we consider as the "reference" power, we fit models (1) and (9) with unknown variances which are usually assumed in practice, although all the power functions are derived under known variance components. We generated 1000 simulated data sets for each combination of pre-specified design parameters and estimated the empirical power as follows:

$$\tilde{\varphi}=\sum_{s=1}^{1000} 1\left\{p_s(\delta)<\alpha\right\}/1000 \qquad (16)$$

where $p_s(\delta)$ is the $p$ value for the $s$th simulated data set ($s$ = 1, 2, …, 1000). The $p$ values were computed based on critical values of Wald $t$-distributions under the null hypotheses with degrees of freedom determined by the method of Kenward and Roger.[20] SAS simulation codes are provided as Supplementary materials.

## Three-level model

The pre-specified design parameters can be found in Table 1. The results show that the theoretical power $\varphi$ (8) and the simulation-based empirical power $\tilde{\varphi}(16)$ are very close to each other regardless of whether $K$ is fixed or varying: $\text{mean}(\varphi)-\text{mean}(\tilde{\varphi})=-0.012$ for fixed $K$ and = 0.001 for varying $K_{ij}$; and $\text{range}(\varphi-\tilde{\varphi})=(-0.049, 0.010)\text{ and }=(-0.041, 0.036)$ respectively. The power function is proven to be an increasing function of all design parameters except $\rho_2$ (3). The effects of $\rho_1$ and $\rho_2$ on the statistical power under three-level model parameters are graphically depicted in Figure 2.

## Two-level model

The pre-specified design parameters can be found in Table 2, in which $\rho_2$ is considered 0. The results show that the performances of the three theoretical power functions $\varphi_{\text{HH}}$ (11), $\varphi_0$ (12), and $\varphi^{(2)}$ (15) are quite different in comparison with the reference empirical power under both fixed $K$ and varying $K_{ij}$: $\text{mean}(\varphi_{\text{HH}})-\text{mean}(\tilde{\varphi})=0.129$ for fixed $K$ and = 0.158 for varying $K_{ij}$; $\text{mean}(\varphi_0)-\text{mean}(\tilde{\varphi})=-0.127$ and = −0.098, respectively; and $\text{mean}(\varphi^{(2)})-\text{mean}(\tilde{\varphi})=0.001$ and = 0.030, respectively. With respect to the ranges of biases: range $(\varphi_{\text{HH}}-\tilde{\varphi})=(0.000, 0.355)$ for fixed $K$ and = (0.018, 0.385) for varying $K_{ij}$, $\text{range}(\varphi_0-\tilde{\varphi})=(-0.183, -0.064)\text{ and }(-0.147, -0.034)$, respectively; and $\text{range}(\varphi^{(2)}-\tilde{\varphi})=(-0.014, 0.021)\text{ and }(0.002, 0.067)$, respectively. Overall, $\varphi^{(2)}$ is least biased and very close to the empirical power. In contrast, $\varphi_{\text{HH}}$ and $\varphi_0$ overestimated and underestimated the empirical power, respectively. Therefore, if a sample size were determined based on $\varphi_{\text{HH}}$ or $\varphi_0$, a study would be under-powered or over-powered, respectively. Furthermore, unlike the other two power functions, $\varphi_{\text{HH}}$ is seen to be increasing with increasing $\rho$ (10) for the values considered for the simulations. For this reason, $\varphi_{\text{HH}}$

could more severely overestimate true power and thus underestimate sample sizes for larger values of $\rho$. The effect of $\rho$ on the statistical power under a two-level model parameters are graphically depicted in Figure 3.

## 7 Discussion

Our results suggest that the second level correlations $\rho_2$ must be accounted for determining sample size when designing a SW assuming a three-level model. However, no SW trials have so far reported an estimate of $\rho_2$, although a couple of SW trial studies[21,22] reported only $\rho_1$ based on the recent review of Davey et al.[23] As observed in this paper (Figure 2), the effects of both $\rho_1$ and $\rho_2$ on the power are substantial when a three-level model is considered. Therefore, it would be valuable to report estimates of $\rho_2$ from conducted SW trials for aiding designs of future SW trials. For two-level models, many studies addressed impacts of $\rho$ (e.g. see literature[12,18,24,25]) as reflected in Figure 3. However, relationship between $\rho$ and $\varphi_{\mathrm{HH}}$ is hardly predictable and mostly contradictory to that between $\rho$ and $\varphi_0$ and that between $\rho$ and $\varphi^{(2)}$ as well.

The derived power function $\varphi$ (8) is proven to be unbiased and valid for that purpose of accounting for both the first and second level correlations. This finding suggests that the pooled estimate $\tilde{\delta}(6)$ may indeed be a maximum likelihood estimate of $\delta$ in model (1). Although it was derived under a special case depicted in Figure 1, the power function $\varphi$ is also proven to be applicable to SW designs with varying $K_{ij}$, the cell size. Therefore, the pooling estimation approach (6) based on the cluster-specific moment estimates can also be extended to the general cases where c varies over steps, and $p$ or $J$ varies over clusters. Nevertheless, even if it could be possible to derive, a closed-form expression of a power function under those situations would be much more complex and much less tractable for calculations.

When $\rho_2$ does not need to be considered in a two-level model, the power function $\varphi^{(2)}$ (15) based on the marginal estimate $\tilde{\delta}_M\,(14)$ performs the best with the ignorable biases regardless of whether the cell sizes are fixed or varying. In contrast, the power function $\varphi_0$ (12) that is reduced from $\varphi$ (8) by plugging 0 into $\rho_2$ in $\varphi$ underestimates the reference power estimated by simulations. This may be because when stratification is unnecessary, pooled estimates can have an unduly inflated variance and thus lose efficiency compared to the marginal approach. On the contrary, the widely used power function $\varphi_{\mathrm{HH}}$ (11) overestimates the reference empirical power and thus underestimates sample sizes under the values of $\rho$ in Table 2. We suspect that the Hussey and Hughes' approach might unduly over or underestimate the variance of the estimate of $\delta$ depending heavily on the values of $\rho$ (Figure 3).

Both models (1) and (9) assume that participants are different across the periods within clusters, let alone between clusters. However, when participants are followed up longitudinally over the periods within the same clusters and crossed over from control to experimental arm, another level of data structure should be modeled by expanding the three-level model (1) to a four-level model that additionally incorporates correlations of outcomes over periods within the same participants. In addition, the random intercepts could be correlated each other violating the independence assumption we took in this paper. Derivations of power functions under these situations design would be a worthy contribution to power literature.

Although only continuous outcome is considered in this paper, categorical or non-normal outcomes such as proportions, incidence rates, ordinal, and survival outcomes are more often of interest in many SW trials.[8] Extension of sample size determinations for such SW trials would be of great interest. The extension might be possible by modeling those outcomes with generalized estimating equations or non-linear mixed-effects models although derivations of closed forms could be intractable. For this reason, sample size determinations based on simulation approaches might be preferable as attempted by Baio et al.[11] Nonetheless, we suspect that simulation of non-normal data with multi-level data hierarchy for a specified correlation structure would be challenging particularly because correlations may well vary with means on which variances depend unlike normal distributions. Therefore, it would also be interesting to examine if extensions based on normal approximations would be comparable. For example, although it has not been verified by simulation studies for a binary outcome, a simple replacement of Δ by $\Delta_p = (p_1 - p_0)/\sqrt{\overline{p}(1-\overline{p})}$ in $\varphi$ or $\varphi^{(2)}$ might be a good approximation owing to a central limit theorem, where $p_0$ and $p_1$ are the "success" probabilities under the null and the alternative hypotheses, respectively, and $\overline{p} = (p_1 + p_0)/2$.

In conclusion, the power functions $\varphi$ (8) and $\varphi^{(2)}$ (15) should be used for sample size determinations for designing SW trials depending on whether the second level correlations $\rho_2$ is assumed to be 0 or not. Both are applicable when cell sizes vary.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was in part supported by the following NIH Grants: UL1RR025750, R01HS023608, and R01DK097096.

## References

1. Hayes, RJ., Moulton, LH. Cluster randomized trials. Boca Raton: CRC Press; 2009.
2. Zhan Z, van den Heuvel ER, Doornbos PM, et al. Strengths and weaknesses of a stepped wedge cluster randomized design: its application in a colorectal cancer follow-up study. J Clin Epidemiol. 2014; 67:454–461. [PubMed: 24491793]
3. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. Clin Trials. 2007; 4:190–199. [PubMed: 17456522]
4. Prost A, Binik A, Abubakar I, et al. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies. Trials. 2015; 16:11. [PubMed: 25560779]

5. Hargreaves JR, Copas AJ, Beard E, et al. Five questions to consider before conducting a stepped wedge trial. Trials. 2015; 16:4. [PubMed: 25558975]
6. Mdege ND, Man M-S, Taylor CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol. 2011; 64:936–948. [PubMed: 21411284]
7. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006; 6:54. [PubMed: 17092344]
8. Beard E, Lewis JJ, Copas A, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. Trials. 2015; 16:14. [PubMed: 25928620]
9. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007; 28:182–191. [PubMed: 16829207]
10. Woertman W, de Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. J Clin Epidemiol. 2013; 66:752–758. [PubMed: 23523551]
11. Baio G, Copas A, Ambler G, et al. Sample size calculation for a stepped wedge trial. Trials. 2015; 16:15. [PubMed: 25592642]
12. Hemming K, Girling A. The efficiency of stepped wedge vs. cluster randomized trials: stepped wedge studies do not always require a smaller sample size. J Clin Epidemiol. 2013; 66:1427–1428. [PubMed: 24035495]
13. Van den Heuvel ER, Zwanenburg RJ, Van Ravenswaaij-Arts CM. A stepped wedge design for testing an effect of intranasal insulin on cognitive development of children with Phelan-McDermid syndrome: a comparison of different designs. Stat Methods Med Res. 2014; doi: 10.1177/0962280214558864
14. Hedges, L., Olkin, I. Statistical methods for meta-analysis. San Diego, CA: Academic Press; 1985.
15. Fazzari MJ, Kim MY, Heo M. Sample size determination for three-level randomized clinical trials with randomization at the first or second level. J Biopharm Stat. 2014; 24:579–599. [PubMed: 24697506]
16. Moerbeek M, van Breukelen GJP, Berger MPF. Design issues for experiments in multilevel populations. J Educ Behav Stat. 2000; 25:271–284.
17. Cohen, J. Statistical power analysis for the behavioral science. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
18. Donner, A., Klar, N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
19. Ahn, C., Heo, M., Zhang, S. Sample size calculations for clustered and longitudinal outcomes in clinical research. Boca Raton, FL: CRC Press; 2014.
20. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997; 53:983–997. [PubMed: 9333350]
21. Bashour HN, Kanaan M, Kharouf MH, et al. 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. 2013; 3:11.
22. Durovni B, Saraceni V, Moulton LH, et al. Effect of improved tuberculosis screening and isoniazid preventive therapy on incidence of tuberculosis and death in patients with HIV in clinics in Rio de Janeiro, Brazil: a stepped wedge, cluster-randomised trial. Lancet Infect Dis. 2013; 13:852–858. [PubMed: 23954450]
23. Davey C, Hargreaves J, Thompson JA, et al. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials. 2015; 16:13. [PubMed: 25588907]
24. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clin Trials. 2005; 2:152–162. [PubMed: 16279137]
25. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. Clin Trials. 2005; 2:99–107. [PubMed: 16279131]

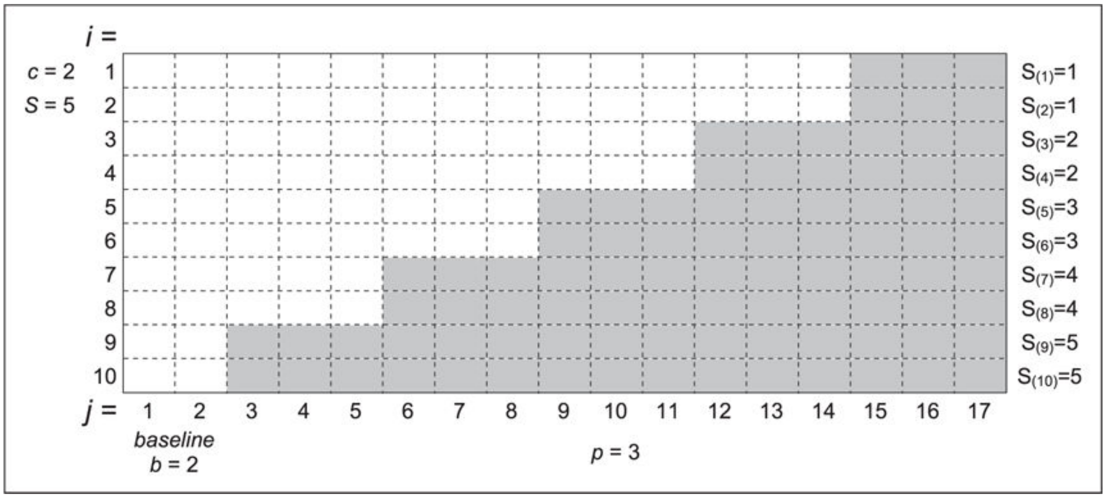

**Figure 1.**

A stepped wedge design with design parameters. Note: Gray areas represent periods under an experimental condition whereas blanks areas represent those under a control condition; $S$ = total number of steps (=5); $c$ = number of clusters per step (=2); $p$ = number of periods per step (=3); $b$ = number of periods at baseline (=2); $I$ = cS = total number of clusters (= 10); $J = b + pS$ = total number of periods per cluster (=17); $N = IJK = cS(b + pS)K$ total number of periods per cluster (=17); $N = IJK = cS(b + pS)K$ =total number of participants (=850) if $K$ = 5, the number of participants per period per cluster.

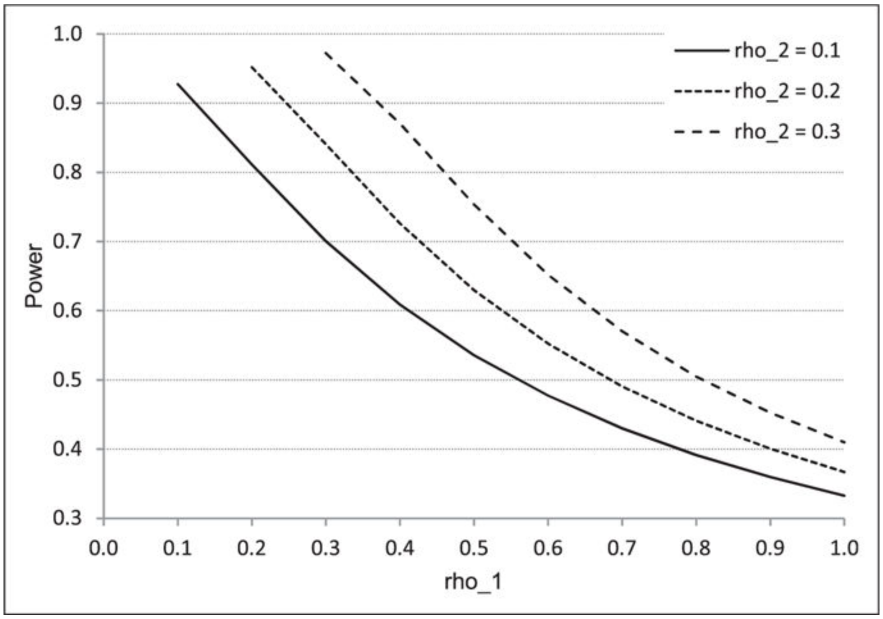

**Figure 2.**

Relationship of $\rho_1$ and $\rho_2$ with statistical power $\varphi$ (8) for a three-level model: Δ = 0.3, $b = 2$, $c = 2$, $p = 2$, $S = 5$, K = 5, and $\alpha = 0.05$. *Note:* rho_1 = $\rho_1$ and rho_2 = $\rho_2$.

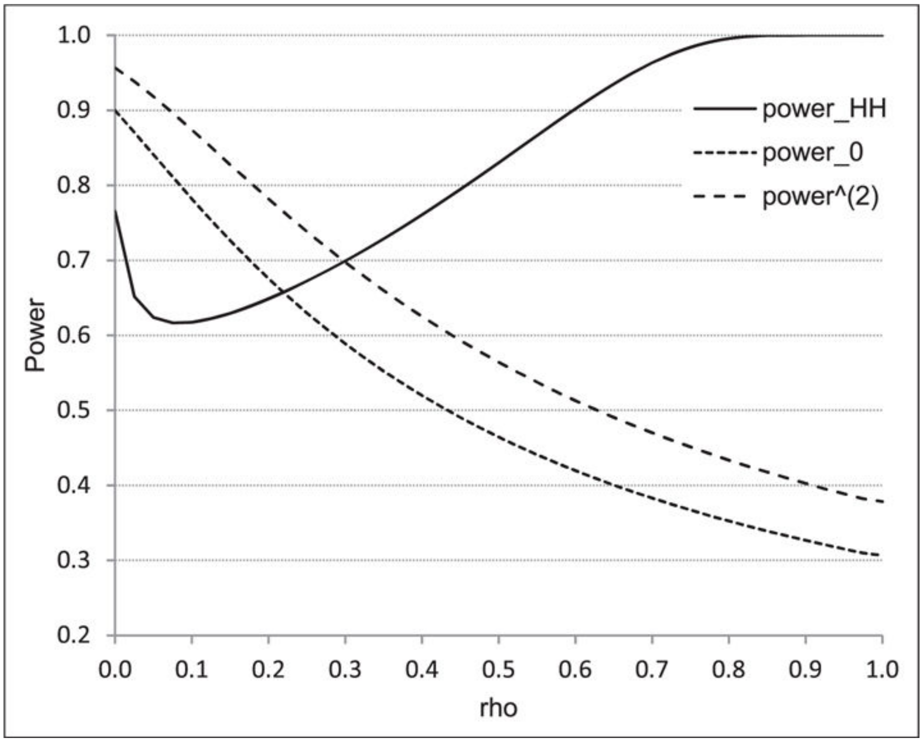

**Figure 3.**

Relationship of $\rho$ with statistical power $\varphi_{HH}$ (11), $\varphi_0$ (12), and $\varphi^{(2)}$ (15) for a two-level model: Δ = 0.3, $b$ = 2, $c$ = 2, $p$ = 2, $S$ = 5, K = 5, and $\alpha$ = 0.05. *Note*: power_HH = $\varphi_{HH}$, power_0 = $\varphi_0$, and powe$\hat{r}$(2) = $\varphi^{(2)}$.

## Table 1

Comparison of theoretical and empirical power with the following design parameters held fixed: $c = 2$, $S = 5$ (or $I = 10$), $p = 2$, and $\sigma^2 = 1$.

| **Δ** | ***b*** | ***J*** | ***K*** | ***N*=*IJK*** | $\rho_1$ | $\rho_2$ | $\varphi$ | $\tilde{\varphi}a$ | ***K*~U(a, b)** | $\tilde{\varphi}b$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 | 0 | 10 | 5 | 500 | 0.3 | 0.1 | 0.539 | 0.559 | U(2, 8) | 0.548 |
| | | | | | | 0.2 | 0.688 | 0.698 | | 0.652 |
| | | | | | 0.4 | 0.1 | 0.457 | 0.500 | | 0.479 |
| | | | | | | 0.2 | 0.564 | 0.567 | | 0.543 |
| | | | 10 | 1000 | 0.3 | 0.1 | 0.637 | 0.655 | U(3, 17) | 0.643 |
| | | | | | | 0.2 | 0.829 | 0.839 | | 0.804 |
| | | | | | 0.4 | 0.1 | 0.516 | 0.540 | | 0.557 |
| | | | | | | 0.2 | 0.653 | 0.647 | | 0.651 |
| | 2 | 12 | 5 | 600 | 0.3 | 0.1 | 0.700 | 0.697 | U(2, 8) | 0.685 |
| | | | | | | 0.2 | 0.841 | 0.836 | | 0.817 |
| | | | | | 0.4 | 0.1 | 0.609 | 0.641 | | 0.620 |
| | | | | | | 0.2 | 0.726 | 0.753 | | 0.723 |
| | | | 10 | 1200 | 0.3 | 0.1 | 0.796 | 0.824 | U(3, 17) | 0.800 |
| | | | | | | 0.2 | 0.940 | 0.931 | | 0.920 |
| | | | | | 0.4 | 0.1 | 0.676 | 0.682 | | 0.678 |
| | | | | | | 0.2 | 0.811 | 0.815 | | 0.802 |
| 0.4 | 0 | 10 | 5 | 500 | 0.3 | 0.1 | 0.783 | 0.802 | U(2, 8) | 0.813 |
| | | | | | | 0.2 | 0.904 | 0.911 | | 0.903 |
| | | | | | 0.4 | 0.1 | 0.695 | 0.730 | | 0.712 |
| | | | | | | 0.2 | 0.807 | 0.821 | | 0.777 |
| | | | 10 | 1000 | 0.3 | 0.1 | 0.868 | 0.858 | U(3, 17) | 0.875 |
| | | | | | | 0.2 | 0.973 | 0.983 | | 0.972 |
| | | | | | 0.4 | 0.1 | 0.760 | 0.809 | | 0.792 |
| | | | | | | 0.2 | 0.881 | 0.903 | | 0.902 |
| | 2 | 12 | 5 | 600 | 0.3 | 0.1 | 0.912 | 0.913 | U(2, 8) | 0.910 |
| | | | | | | 0.2 | 0.976 | 0.974 | | 0.971 |
| | | | | | 0.4 | 0.1 | 0.846 | 0.853 | | 0.842 |
| | | | | | | 0.2 | 0.927 | 0.930 | | 0.921 |

| **Δ** | ***b*** | ***J*** | ***K*** | ***N*=*IJK*** | $\boldsymbol{\rho_1}$ | $\boldsymbol{\rho_2}$ | $\boldsymbol{\varphi}$ | $\tilde{\varphi}$[a] | ***K*~U(a, b)** | $\tilde{\varphi}$[b] |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 10 | 1200 | 0.3 | 0.1 | 0.961 | 0.970 | U(3, 17) | 0.949 |
| | | | | | | 0.2 | 0.997 | 0.994 | | 0.995 |
| | | | | | 0.4 | 0.1 | 0.896 | 0.909 | | 0.903 |
| | | | | | | 0.2 | 0.966 | 0.963 | | 0.955 |
| Mean | | | | | | | 0.785 | 0.797 | | 0.785 |

[a] Empirical power under a fixed *K*.

[b] Empirical power under a varying *K* ~ U(a,b) following a uniform distribution with mean equal to the corresponding fixed *K*.

### Table 2

Comparison of theoretical and empirical power with the following design parameters held fixed: $c = 2$, $S = 5$ (or $I = 10$), $p = 2$, and $\sigma^2 = 1$.

| **Δ** | ***b*** | ***J*** | ***K*** | ***N=IJK*** | ***ρ*** | $\varphi_{HH}$ | $\varphi_0$ | $\varphi^{(2)}$ | $\tilde{\varphi}$[a] | ***K~U*(a, b)** | $\tilde{\varphi}$[b] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.3 | 0 | 10 | 5 | 500 | 0.3 | 0.640 | 0.440 | 0.600 | 0.579 | U(2, 8) | 0.565 |
| | | | | | 0.4 | 0.700 | 0.384 | 0.531 | 0.532 | | 0.464 |
| | | | 10 | 1000 | 0.3 | 0.900 | 0.505 | 0.676 | 0.675 | U(3, 17) | 0.642 |
| | | | | | 0.4 | 0.937 | 0.424 | 0.582 | 0.582 | | 0.552 |
| | 2 | 12 | 5 | 600 | 0.3 | 0.699 | 0.589 | 0.697 | 0.696 | U(2, 8) | 0.653 |
| | | | | | 0.4 | 0.761 | 0.520 | 0.625 | 0.626 | | 0.623 |
| | | | 10 | 1200 | 0.3 | 0.937 | 0.664 | 0.771 | 0.770 | U(3, 17) | 0.738 |
| | | | | | 0.4 | 0.964 | 0.570 | 0.678 | 0.682 | | 0.642 |
| 0.4 | 0 | 10 | 5 | 500 | 0.3 | 0.871 | 0.674 | 0.840 | 0.850 | U(2, 8) | 0.802 |
| | | | | | 0.4 | 0.912 | 0.602 | 0.776 | 0.775 | | 0.749 |
| | | | 10 | 1000 | 0.3 | 0.991 | 0.749 | 0.896 | 0.901 | U(3, 17) | 0.879 |
| | | | | | 0.4 | 0.996 | 0.655 | 0.824 | 0.838 | | 0.793 |
| | 2 | 12 | 5 | 600 | 0.3 | 0.911 | 0.830 | 0.910 | 0.911 | U(2, 8) | 0.893 |
| | | | | | 0.4 | 0.945 | 0.764 | 0.859 | 0.850 | | 0.835 |
| | | | 10 | 1200 | 0.3 | 0.996 | 0.888 | 0.950 | 0.952 | U(3, 17) | 0.922 |
| | | | | | 0.4 | 0.999 | 0.813 | 0.898 | 0.884 | | 0.884 |
| Mean | | | | | | 0.885 | 0.629 | 0.757 | 0.756 | | 0.727 |

[a] Empirical power under a fixed $K$.

[b] Empirical power under a varying $K \sim$ U(a,b) following a uniform distribution with mean equal to the corresponding fixed $K$.