Article

Journal of Educational and Behavioral Statistics
2018, Vol. 43, No. 2, pp. 159–181
DOI: 10.3102/1076998617730303

The Design of Cluster Randomized Trials With Random Cross-Classifications

Mirjam Moerbeek
Utrecht University

Maryam Safarkhani
MSD

Data from cluster randomized trials do not always have a pure hierarchical structure. For instance, students are nested within schools that may be crossed by neighborhoods, and soldiers are nested within army units that may be crossed by mental health–care professionals. It is important that the random cross-classification is taken into account while planning a cluster randomized trial. This article presents sample size equations, such that a desired power level is achieved for the test on treatment effect. Furthermore, it also presents optimal sample sizes given a budgetary constraint, with a special focus on conditional optimal designs where one of the sample sizes is fixed beforehand. The optimal design methodology is illustrated using a postdeployment training to reduce ill-health in armed forces personnel.

Keywords: *cluster randomization; crossed random effects; statistical power analysis; optimal design; budgetary constraint; conditional optimal design*

Cluster randomized trials are conducted to evaluate the effectiveness of interventions in the behavioral, health, and biomedical sciences. Examples are health promotion interventions to improve vaccine uptake with students nested within schools and psychotherapy interventions to reduce post-traumatic stress disorder with soldiers nested within military units such as platoons or companies. The basic feature of this type of trial is that subjects are nested within clusters and randomization is done at the cluster level, which implies all subjects within the same cluster receive the same treatment. Although cluster randomization is less efficient than individual-level randomization, it is often chosen for political, financial, and ethical reasons and to avoid the risk of contamination of the control condition (Gail, Mark, Carroll, & Green, 1996; Moerbeek, 2005). For a basic introduction to the cluster randomized design, the reader is referred to Campbell and Walters (2014), Donner and Klar (2000), Eldridge and Kerry (2012), Hayes and Moulton (2009), and Murray (1998).

Measurements of outcomes from subjects within the same cluster are likely to be correlated because of mutual influence and shared norms. Such correlations should be taken into account while analyzing the data, and in this article, the multilevel regression model is used for that purpose (Goldstein, 2011; Hox, 2010; Snijders & Bosker, 2012). It is important that all levels of nesting are identified and included in the multilevel model. Ignoring one or more levels may result in biased estimates of treatment effects and, in particular, their standard errors, which may result in inflated type I or type II error rates (Moerbeek, 2004; Moerbeek, Van Breukelen, & Berger, 2003a; Opdenakker & Van Damme, 2000). In a similar manner, the results of cluster randomized trials may also be biased when imperfect hierarchies, such as random cross-classifications, are present in the data but ignored in the data analysis (Gilbert, Petscher, Compton, & Schatschneider, 2016; Luo & Kwok, 2009; Meyers & Beretvas, 2006). An example of a cross-classified design is a cluster randomized trial where soldiers are nested within army units and treated by mental health–care professionals (Castro, Adler, McGurk, & Bliese, 2012; Mulligan et al., 2012). As not all soldiers from the same unit are treated by the same professional, units and professionals are crossed with each other. Cross-classified structures can easily be taken into account by including a random effect for each crossed factor (Goldstein, 1994; Rasbash & Goldstein, 1994), such as for the units and professionals in the example. This implies a variance component is estimated for each random factor to represent the between-unit and between-professional variance in the outcome variable. The higher these variances, the more the units and the more the professionals differ from each other with respect to the outcome variable and the more severe the consequences of ignoring crossed random factors while analyzing data from a cluster randomized trial.

The possibility of random cross-classifications should not only be considered while analyzing data but also while designing cluster randomized trials. Part of a good design is a proper sample size calculation: How many clusters and which cluster size are needed to detect a treatment effect with a desired probability? Over the past two decades, much attention has been paid to sample size calculations for cluster randomized trials with two levels of nesting (Raudenbush, 1997), three levels of nesting (Cunningham & Johnson, 2012; Heo & Leon, 2008; Konstantopoulos, 2009; Moerbeek, Van Breukelen, & Berger, 2000; Teerenstra, Moerbeek, Van Achterberg, Pelzer, & Borm, 2008), dichotomous outcomes (Moerbeek, Van Breukelen, & Berger, 2001), and 2×2 factorial designs (Lemme, Van Breukelen, & Berger, 2015; Moerbeek, Van Breukelen, & Berger, 2003b). Thus far, attention has been restricted to cluster randomized trials with a pure hierarchical data structure. In terms of the example, sample size equations are available for trials where soldiers are nested within units, but the nesting within professionals is ignored. In other words, such sample size equations inform us how many units are needed and how many soldiers per unit should be sampled, but they do not inform us how many professionals should be

included in the trial and how many soldiers each of them should treat. This may result in a design that is suboptimal and therefore a misuse of the researchers' and participants' time and efforts.

The aim of this contribution is to present sample size equations for cluster randomized trials with random cross-classifications. It is important to distinguish two different types of random cross-classifications in cluster randomized trials: those that exist naturally and those that do not. The cross-classification of army units and health-care professionals is an example of the latter. A pure hierarchy with soldiers nested within units exists before randomization of units to treatment conditions, and the random cross-classification is established once soldiers are assigned to professionals. In a completely crossed design, data are available in each cell (i.e., in each unit-professional combination). In the literature of trials of therapist interventions, this design has been referred to as a cross-therapist design. One can also image a design where the first set of units is treated by the first handful of professionals, the second set of units is treated by the second handful of professionals, and so on. This is a partially crossed design (or a nested-therapist design) and may be encountered when not all units return from combat at the same time and not all professionals are available for the whole duration of the study. By clever randomization of soldiers to professionals, one can achieve a balanced design. Balanced designs are favored, as they are more efficient than unbalanced designs (Van Breukelen & Candel, 2012; Van Breukelen, Candel, & Berger, 2007).

An example of a cluster randomized trial with a naturally existing random cross-classification is an intervention study to improve vaccine uptake (Ali et al., 2007) with students nested within schools and crossed by neighborhoods. The number of neighborhoods served by schools is likely to be highly variable, and neighborhoods serving different schools will only be partially overlapping at the best. In such an example, the design is likely to be unbalanced. One may try to achieve a balanced design by subsampling students from each school–neighborhood combination, but this may not always be justified in practice. For instance, it may not be considered ethical to offer an intervention to some students in a school and to refrain other students in the same school from that intervention. In the most extreme case, some cells may not have any students at all, which implies a balanced design cannot be achieved by subsampling.

The organization of this article is as follows. In the next section, the multilevel model for cross-classified data is presented, and the partitioning of the variance to the different levels is given. The third section studies statistical power and optimal design for the complete random cross-classification. It presents mathematical expressions for the variance of the treatment effect estimator, as this variance will be used to calculate the required sample size to achieve a desired power level for the tests on main and interaction effects. It will be shown that the power level depends on the number of units, the number of professionals, and the number of soldiers per cell. Once two of them are fixed, the third can be

calculated such that the desired power level is achieved. When neither sample size is fixed beforehand, the best combination of these three sample sizes can be determined by taking a budgetary constraint into account. Special attention is also paid to conditional optimal designs (Hedges & Borenstein, 2014), where one of the three sample sizes is fixed and the cost constraint is used to calculate the other two. The focus of fourth section is on partial random cross-classifications and the loss of efficiency as compared to the complete random cross-classification. The third and fourth sections restrict to balanced designs; an extension to unbalanced designs is made in the fifth section. The final section gives a summary of the results, a discussion, and directions for future research.

Statistical Model and Variance Partitioning

Army units are at random assigned such that half of them receive the intervention and the others receive the control. Within each unit, soldiers are subsequently assigned to mental health–care professionals. The regression model relates outcome y_{ijk} of soldier i in unit j and professional k to treatment condition x_j:

$$y_{ijk} = \gamma_0 + \gamma_1 x_j + u_j + v_k + w_{jk} + e_{ijk}. \tag{1}$$

Treatment condition x_j is coded by a dummy variable that takes on the value -0.5 for the control and $+0.5$ for the intervention, which implies the slope γ_1 is the treatment effect size. It should be mentioned the effect of treatment is constant across the therapists. Equation (1) contains four random effects: $u_j \sim N(0, \sigma_u^2)$ is the random effect of unit j, $v_k \sim N(0, \sigma_v^2)$ that of professional k, $w_{jk} \sim N(0, \sigma_w^2)$ that of the interaction between unit j and professional k, and $e_{ijk} \sim N(0, \sigma_e^2)$ that of soldier i in unit j and professional k. It is important to include the possibility of an interaction between the random effects of unit and professional (Shi, Leite, & Algina, 2010). The random effects are assumed to be independent of each other, which implies the total variance is $\text{var}(y_{ijk}) = \sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2$. As the variance is a constant, the model assumes homoscedasticity. The covariance of outcomes of two students depends on whether they come from the same unit and/or are treated by the same professional. The covariance of two soldiers who come from the same unit j and are treated by the same professional k is $\text{cov}(y_{ijk}, y_{ijk}) = \sigma_u^2 + \sigma_v^2 + \sigma_w^2$; the covariance of two soldiers from the same unit j who are treated by different professionals k and k' is $\text{cov}(y_{ijk}, y_{ijk'}) = \sigma_u^2$, and the covariance of soldiers who are treated by the same professional k but come from different units j and j' is $\text{cov}(y_{ijk}, y_{ij'k}) = \sigma_v^2$. It is obvious the covariance is higher when the two soldiers come from the same unit and professional.

The intraclass correlations quantify the proportion variance that is attributable to each random effect. The proportion variance at the level of the unit is

$$\rho_u = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2}. \tag{2}$$

Similarly, the intraclass correlation at the level of the professional is

$$\rho_v = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2}, \tag{3}$$

and the intraclass correlation due to the random unit by professional interaction is

$$\rho_w = \frac{\sigma_w^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2}. \tag{4}$$

These proportions vary between 0 and 1. The proportion variance at the soldier level is $1 - \rho_u - \rho_v - \rho_w$.

Complete Random Cross-Classification

In the design with a complete random cross-classification, data are available in each cell. This implies each health professional delivers both the intervention and control. The sample size calculations that follow in this section assume a balanced design: Randomization of soldiers to professionals is done such that each cell has the same amount of soldiers n_1. The number of units is denoted by n_{2A}, and the number of professionals is denoted by n_{2B}, which implies the total sample size is $n_1 n_{2A} n_{2B}$.

Given a balanced design, the treatment effect size γ_1 is estimated by the difference in mean scores in the intervention and control conditions, and this estimator has variance

$$\begin{aligned} \text{var}(\hat{\gamma}_1) &= \frac{4(\sigma_e^2 + n_1 n_{2B}\sigma_u^2 + n_1\sigma_w^2)}{n_1 n_{2A} n_{2B}} \\ &= \frac{4(\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2)(1 - \rho_v + (n_1 n_{2B} - 1)\rho_u + (n_1 - 1)\rho_w)}{n_1 n_{2A} n_{2B}}. \end{aligned} \tag{5}$$

This variance is estimated by replacing the variance components by their estimates. A justification of Equation 5 is given in Appendix 1 of the Supplemental Material.

The variance for the treatment effect estimator depends on the variance component σ_u^2 for units but not on the variance component σ_v^2 for professionals. This is obvious since the treatment condition varies between units and within professionals. This implies a higher power to detect the effect of treatment when the between health professional variation increases, assuming the total variance remains constant. However, this assumes that there is no variation in the treatment effect between health professionals that would be modeled by a health professional random coefficient for treatment (see, e.g., Walwyn & Roberts,

2010). If this random effect is present, power might be reduced by the random coefficient for treatment.

Note that the unit variance component σ_u^2 is multiplied by the total number of soldiers $n_1 n_{2B}$ within each unit. In other words, the design becomes less efficient when there are many soldiers within each unit and/or the between-unit variance is large. In addition, the efficiency decreases with increasing number of soldiers per cell (n_1) and with increasing covariance σ_w^2 between units and professionals. The number of units n_{2A} is the only sample size that does not appear in the numerator of Equation 5, so the effect of increasing n_{2A} on efficiency is much stronger than the effects of increasing n_1 and n_{2B}.

The term $(1 - \rho_v + (n_1 n_{2B} - 1)\rho_u + (n_1 - 1)\rho_w)$ in the numerator at the right side is called the design effect. It is the factor with which the variance of the treatment effect estimator in an individually randomized trial should be multiplied to get the variance in a cluster randomized trial with a random cross-classification. This factor is larger than 1 when $\rho_u(n_1 n_{2B} - 1) + \rho_w(n_1 - 1) > \rho_v$, a condition which is likely to be fulfilled.

The Relation Between Sample Size and Power

Equation 5 shows that the precision with which the effect of intervention is estimated increases with increasing sample size. In other words, a larger sample size implies a higher statistical power to detect the intervention effect, provided it exists in the population. In trials like this, there is not just one sample size but three: the number of units n_{2A}, the number of professionals n_{2B}, and the number of soldiers per cell n_1. Trials with the same total sample size may result in different power levels: Equation 5 shows the variance of the estimator does not only depend on the total sample size but also on the number of professionals and the number of soldiers per cell.

The test statistic for the test on intervention effect is calculated as $t = \hat{\gamma}_1 / \widehat{s.e.}(\hat{\gamma}_1)$, and under the null hypothesis of no effect ($H_0 : \gamma_1 = 0$), it follows a central t distribution with degrees of freedom equal to $df = n_{2B} - 1$. Let us assume the alternative hypothesis is one-sided, and higher scores are expected in the intervention condition: $H_A : \gamma_1 > 0$. Under the alternative hypothesis, the test statistic follows a noncentral t distribution with the same degrees of freedom and noncentrality parameter $\lambda = \gamma_1 / \sqrt{\left(4(\sigma_e^2 + n_1 n_{2B}\sigma_u^2 + n_1\sigma_w^2)/n_1 n_{2A} n_{2B}\right)}$. The null hypothesis is rejected when the test statistic exceeds the critical value $t_{n_{2B}-1,1-\alpha}$, which is the $100(1-\alpha)$th percentile of the central t distribution with $n_{2B} - 1$ degrees of freedom and α the type I error rate. The power of the test is calculated as

$$1 - \beta = P(t_{n_{2B}-1}(\lambda) > t_{n_{2B}-1,1-\alpha}), \tag{6}$$

where the random variable $t_{df}(\lambda)$ follows a noncentral t distribution with noncentrality parameter λ and df degrees of freedom. For the one-sided alternative hypothesis $H_A : \gamma_1 < 0$ (i.e., higher outcomes expected in the control condition), the power follows from

$$1 - \beta = P(t_{n_{2B}-1}(\lambda) < t_{n_{2B}-1,\alpha}), \tag{7}$$

and for the two-sided alternative hypothesis $H_A : \gamma_1 \neq 0$, it follows from

$$1 - \beta = P(t_{n_{2B}-1}(\lambda) > t_{n_{2B}-1,1-\alpha/2}) + P(t_{n_{2B}-1}(\lambda/2) < t_{n_{2B}-1,\alpha/2}). \tag{8}$$

For large degrees of freedom, the t distribution can be approximated by the standard normal and the general equation for the relation between sample sizes and power for the test on the intervention effect γ_1 is

$$\text{var}(\gamma_1) = \left(\frac{\gamma_1}{z_{1-\alpha} + z_{1-\beta}}\right)^2, \tag{9}$$

where $z_{1-\alpha}$ and $z_{1-\beta}$ are the $100(1-\alpha)$th and $100(1-\beta)$th percentiles from the standard normal distribution. This equation holds for tests with a one-sided alternative hypothesis; for a two-sided alternative, $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$.

In many trials, some of the sample sizes n_1, n_{2A}, and n_{2B} are fixed beforehand. For instance, the number of units that are willing to participate in the intervention may be limited or the number of soldiers that can be included from each unit may be limited by unit size. When two sample sizes are fixed, the third can be calculated such that the desired power level $1 - \beta$ is achieved, and these sample sizes are given in Table 1.

It is obvious the required sample size increases when the other two sample sizes decrease. In addition to that, a larger sample size is needed when smaller type I and type II error rates α and β are required. Furthermore, sample size depends on the size of the effect γ_1 and the variance components σ_e^2, σ_u^2, and σ_w^2. The size of the effect is often unknown while designing an experiment, and an educated guess has to be made on the basis of a prior study, the literature, or an expert opinion. Alternatively, the minimally relevant effect size may be used. The effect size is often expressed as a standardized effect size, which is the difference in mean scores in both treatment conditions divided by the standard deviation of the outcome: $\delta = \gamma_1 / \sqrt{\sigma_e^2 + \sigma_u^2 + \sigma_v^2 + \sigma_w^2}$. Cohen (1988) distinguishes small ($\delta = .2$), medium ($\delta = .5$), and large ($\delta = .8$) effect sizes. An educated guess of the variance components can be based on estimates as published in the literature. Since the mid-1990s, some 60 papers have appeared that published estimates of intraclass correlation coefficients, and an overview is provided by Moerbeek and Teerenstra (2016).

TABLE 1.
Required Number of Units n_{2A}*, Number of Professionals* n_{2B}*, and Number of Soldiers* n_1 *Per Cell When the Other Two Are Known*

Scenario	Required Sample Size
n_1 and n_{2B} fixed, calculate n_{2A}	$n_{2A} = 4\frac{\sigma_e^2 + n_1 n_{2B}\sigma_u^2 + n_1\sigma_w^2}{n_1 n_{2B}} \times \frac{1}{\left(\frac{\gamma_1}{z_{1-\alpha}+z_{1-\beta}}\right)^2}$ $= 4\frac{1-\rho_v+(n_1 n_{2B}-1)\rho_u+(n_1-1)\rho_w}{n_1 n_{2B}} \times \frac{1}{\left(\frac{\delta}{z_{1-\alpha}+z_{1-\beta}}\right)^2}$
n_1 and n_{2A} fixed, calculate n_{2B}	$n_{2B} = 4\frac{\sigma_e^2 + n_1\sigma_w^2}{n_1 n_{2A}} \times \frac{1}{\left(\frac{\gamma_1}{z_{1-\alpha}+z_{1-\beta}}\right)^2 - 4\frac{\sigma_v^2}{n_{2A}}}$ $= 4\frac{1-\rho_v-\rho_u+(n_1-1)\rho_w}{n_1 n_{2A}} \times \frac{1}{\left(\frac{\delta}{z_{1-\alpha}+z_{1-\beta}}\right)^2 - 4\frac{\rho_v}{n_{2A}}}$
n_{2A} and n_{2B} fixed, calculate n_1	$n_1 = 4\frac{\sigma_e^2}{n_{2A}n_{2B}} \times \frac{1}{\left(\frac{\gamma_1}{z_{1-\alpha}+z_{1-\beta}}\right)^2 - 4\frac{\sigma_v^2}{n_{2A}} - 4\frac{\sigma_w^2}{n_{2A}n_{2B}}}$ $= 4\frac{1-\rho_v-\rho_u-\rho_w}{n_{2A}n_{2B}} \times \frac{1}{\left(\frac{\delta}{z_{1-\alpha}+z_{1-\beta}}\right)^2 - 4\frac{\rho_v}{n_{2A}} - 4\frac{\rho_w}{n_{2A}n_{2B}}}$

Optimal Sample Sizes Given a Budgetary Constraint

Here, we focus on the derivation of the optimal combination of the number of units, professionals, and soldiers per cell: $\xi^* = (n_1^*, n_{2B}^*, n_{2A}^*)$. In practice, these sample sizes are often limited by budgetary constraints, and to derive such an optimal combination, a cost function must be defined as a precondition.

Assume the costs to include a unit are c_{2A} currency units, the costs to train a professional are c_{2B} currency units, and that measuring an outcome from a soldier costs c_1 currency units. The costs that are needed for the trial with n_{2B} professionals, n_{2A} units, and n_1 soldiers per cell should not exceed the budget C:

$$c_{2A}n_{2A} + c_{2B}n_{2B} + c_1 n_1 n_{2A} n_{2B} < C. \quad (10)$$

Here, $c_{2A}n_{2A}$ is the amount of the total budget spent on the unit-level costs, $c_{2B}n_{2B}$ is the amount for the professional-level costs, and $c_1 n_1 n_{2A} n_{2B}$ is the amount for the soldier-level costs. In addition to the budgetary constraint, we allow for maximum constraints on the numbers of units and professionals since both may be limited in practical settings.

The next step in finding an optimal design is to choose an optimality criterion, which must be minimized under the above cost constraint. In this article, we consider the variance of the treatment effect estimator, $\text{var}(\hat{\gamma}_1)$, as minimum variance results in maximal power of the test on treatment effect. Finding an optimal design means finding the number of units, the number of professionals, and the number of soldiers per cell that minimize the optimality criterion given the constraint in Equation 10.

The relative efficiency measures how well any other design ξ performs as compared to the optimal design ξ^*. It is defined as

$$\text{RE}^{\xi|\xi^*} = \frac{\text{var}(\hat{\gamma}_1)_{\xi^*}}{\text{var}(\hat{\gamma}_1)_{\xi}},$$

where the numerator is $\text{var}(\hat{\gamma}_1)$ under the optimal design ξ^* and the denominator is the variance for the other design ξ. The $\text{RE}^{\xi|\xi^*}$ lies between 0 and 1, and if $\text{RE}^{\xi|\xi^*} = 1$, the two designs ξ and ξ^* are equally efficient. The $\text{RE}^{\xi|\xi^*}$ can be interpreted in terms of sample size. For example, if $\text{RE}^{\xi|\xi^*} = .8$, then $(0.8^{-1} - 1) \times 100\% = 25\%$, which implies the sample size of design ξ needs to be increased by 25% to achieve the same efficiency as for design ξ^*. Therefore, relative efficiencies of 0.8 or 0.9 and closer to 1 are often preferred.

Unfortunately, there exist no simple closed-form formulas for the optimal sample sizes, and the optimal design should therefore be found on the basis of numerical techniques. R code for this purpose is available from the first author upon request. This R code also calculates the RE of alternative designs as compared to the optimal design. The use of the code will be demonstrated on the basis of an example in An Example: Postdeployment Training to Reduce Ill-Health in Armed Forces Personnel subsection.

Conditional Optimal Designs

In practice, it may occur that the number of army units, the number of health professionals, or the number of soldiers per cell is fixed beforehand. For instance, one may want to include all soldiers within a unit in the intervention and then the budgetary constraint is used to determine the optimal numbers of units and professionals. This is a so-called conditional optimal design, and such designs

TABLE 2.
Conditional Optimal Designs in Cluster Randomized Trials With Random Cross-Classifications

Scenario	Optimal Sample Sizes
n_{2A} fixed to $\tilde{n}_{2A}$	$n_1^* = \sqrt{\frac{\sigma_e^2 c_{2B}}{\sigma_w^2 c_1 \tilde{n}_{2A}}} = \sqrt{\frac{(1-\rho_v-\rho_u-\rho_w)c_{2B}}{\rho_w c_1 \tilde{n}_{2A}}}$
	$n_{2B}^* = \frac{C-c_{2A}\tilde{n}_{2A}}{c_1 n_1^* \tilde{n}_{2A}+c_{2B}} = \frac{C-c_{2A}\tilde{n}_{2A}}{\sqrt{\frac{c_1 \tilde{n}_{2A}\sigma_e^2 c_{2B}}{\sigma_w^2}}+c_{2B}} = \frac{C-c_{2A}\tilde{n}_{2A}}{\sqrt{\frac{c_1 \tilde{n}_{2A}(1-\rho_v-\rho_u-\rho_w)c_{2B}}{\rho_w}}+c_{2B}}$
n_{2B} fixed to $\tilde{n}_{2B}$	$n_1^* = \sqrt{\frac{\sigma_e^2 c_{2A}}{c_1 \tilde{n}_{2B}(\tilde{n}_{2B}\sigma_u^2+\sigma_w^2)}} = \sqrt{\frac{(1-\rho_v-\rho_u-\rho_w)c_{2A}}{c_1 \tilde{n}_{2B}(\tilde{n}_{2B}\rho_u+\rho_w)}}$
	$n_{2A}^* = \frac{C-c_{2B}\tilde{n}_{2B}}{c_1 n_1^* \tilde{n}_{2B}+c_{2A}} = \frac{C-c_{2B}\tilde{n}_{2B}}{\sqrt{\frac{\tilde{n}_{2B}c_1\sigma_e^2 c_{2A}}{(\tilde{n}_{2B}\sigma_u^2+\sigma_w^2)}}+c_{2A}} = \frac{C-c_{2B}\tilde{n}_{2B}}{\sqrt{\frac{\tilde{n}_{2B}c_1(1-\rho_v-\rho_u-\rho_w)c_{2A}}{(\tilde{n}_{2B}\rho_u+\rho_w)}}+c_{2A}}$
n_1 fixed to $\tilde{n}_1$	$n_{2A}^* = \frac{C-c_{2B}n_{2B}^*}{c_1 \tilde{n}_1 n_{2B}^*+c_{2A}}$
	$n_{2B}^* = \frac{\sqrt{1+\frac{Cc_1\tilde{n}_1}{c_{2A}c_{2B}}S}-1}{\frac{c_1\tilde{n}_1}{c_{2A}}S}$, where
	$S = \frac{Cc_1\tilde{n}_1+c_{2A}c_{2B}}{c_1 c_{2B}(\frac{\sigma_e^2+\tilde{n}_1\sigma_w^2}{\sigma_u^2})} + 1 = \frac{Cc_1\tilde{n}_1+c_{2A}c_{2B}}{c_1 c_{2B}(\frac{(1-\rho_v-\rho_u+(\tilde{n}_1-1)\rho_w)}{\rho_u})} + 1$

have been derived for three- and four-level cluster randomized trials without a random cross-classification (Hedges & Borenstein, 2014).

Table 2 shows expressions for conditional optimal designs in cluster randomized trials with a random cross-classification and the budgetary constraint (Equation 10; see Appendix 2 of the Supplemental Material for the derivation of these conditional optimal designs). In the first scenario, n_{2A} is fixed to a constant. The optimal cell size n_1^* depends on the cost ratio c_{2B}/c_1; it makes sense the optimal cell size increases when the professional-level costs c_{2B} increase and the soldier-level costs c_1 decrease. Furthermore, the optimal n_1^* increases with decreasing number of army units $\tilde{n}_{2A}$ since a higher amount of money is available for enrolling soldiers when fewer units are included. Finally, the optimal cell size increases when the between-soldier variance increases and/or the variance of the unit-profession interaction decreases. It should be noted the optimal cell size does not depend on the budget C. The optimal number of professionals n_{2B}^* follows from the cost function; it does depend on the budget and it is inversely related to the optimal cell size.

The optimal sample sizes for the scenario where n_{2B} is fixed are rather similar to the first scenario and not further discussed. The optimal sample sizes for the third scenario where n_1 is fixed are much more complicated and are given in the final two rows of Table 2. Here, both the number of units and the number of professionals are a function of the budget C.

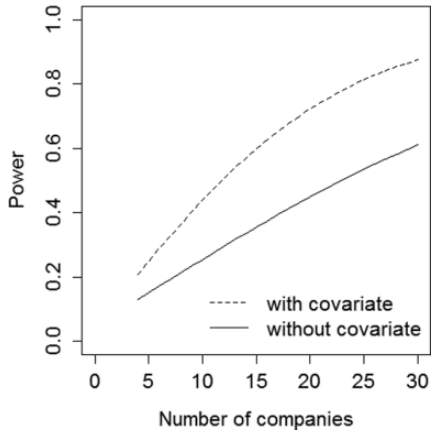

FIGURE 1. *Power as a function of the number of companies* n_{2A} *and for cell size* $n_1 = 8$ *and* $n_{2B} = 12$ *professionals. Intraclass correlation coefficients:* $\rho_u = .3$, $\rho_v = .1$, *and* $\rho_w = .05$.

For all three scenarios, R code is available from the first author upon request and can be used to find the optimal design and to evaluate the relative efficiency of suboptimal designs. Its use will be demonstrated in the next section.

An Example: Postdeployment Training to Reduce Ill-Health in Armed Forces Personnel

In the study by Mulligan et al. (2012), companies of approximate size 100 were randomly assigned to an intervention or control and treated by 12 health-care professionals. The intervention was a postdeployment psychoeducational intervention, and the control was a standard brief. The primary outcomes were the symptoms of post-traumatic stress and common mental disorders. Secondary outcomes were depression, sleep quality, alcohol misuse, and stigmatizing beliefs regarding seeking help for or having a mental health problem. In their power analysis, they assumed a medium standardized effect size and an intraclass correlation at the company level of $\rho_u = .3$. An intraclass correlation at the level of the professional was not given, neither that of the interaction between company and professional. Without further justification, we use $\rho_v = .1$ and $\rho_w = .05$ in our power calculations that follow.

Let us consider a scenario where the number of professionals and the size of the company are fixed, and the number of companies to achieve a power of 80% is calculated. Figure 1 shows the power as a function of the number of companies up to $n_{2A} = 30$, and for $n_{2B} = 12$, professionals and cell size $n_1 = 8$ (i.e., each company has size $n_1 n_{2B} = 96$, so that the design is balanced). A test with a

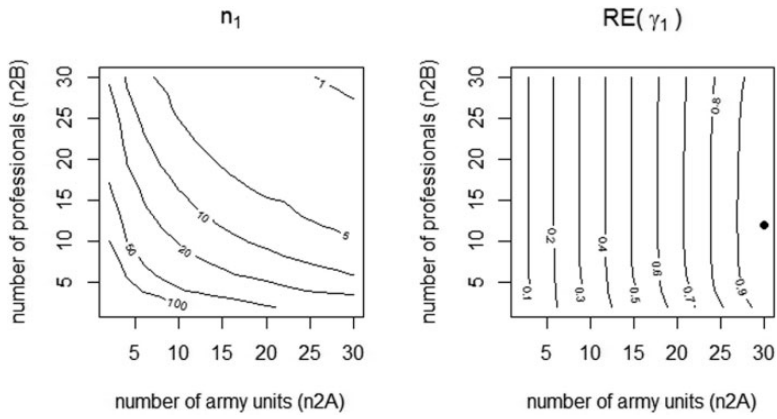

FIGURE 2. *Contour plot for the number of soldiers per cell (left graph) and for the relative efficiency (right graph). The dot represents the optimal design.* $C = 2{,}500$, $c_{2A} = 15$, $c_{2B} = 45$, $\rho_u = .3$, $\rho_v = .1$, *and* $\rho_w = .05$.

one-sided alternative and type I error rate $\alpha = .05$ is used; the standardized effect size is assumed to be medium (i.e., $\delta = .5$). The power achieved with the maximum amount of $n_{2A} = 30$ companies is $1 - \beta = .61$, see the bold line in Figure 1.

As increasing the number of professionals and/or the number of soldiers per cell only has a minor effect on power, other means should be used to get a more acceptable power level of $1 - \beta = .8$. The dashed line in Figure 1 shows the power for the scenario where a predictive covariate at the company level explains 50% of the between company variance σ_u^2. Here, a power of $1 - \beta = .8$ can be achieved with $n_{2A} = 24$ companies. As the between-professional variance σ_v^2 does not contribute to $\text{var}(\hat{\gamma}_1)$, adding covariates at the level of the professional to the equation (1) does not increase power. Similarly, adding covariates at the soldier level to the model only slightly improves power since the error variance σ_e^2 only slightly contributes to $\text{var}(\hat{\gamma}_1)$.

In case neither sample size is fixed beforehand, a budgetary constraint can be used to derive the optimal design. Suppose the maximum number of companies that can be recruited is $n_{2A\max} = 30$ and the maximum number of professionals that is available is $n_{2B\max} = 30$. As such we acknowledge the number of companies and professionals is often limited in practice. To derive the optimal design, prior estimates of the variance components must be given and the costs must be specified. We use the same values for ρ_u, ρ_v, and ρ_w as above and consider a trial with a budget of $C = 2{,}500$, costs per company equal to $c_{2A} = 15$, costs per professional equal to $c_{2B} = 45$, and costs per soldier equal to $c_1 = 1$.

The contour graph at the left of Figure 2 shows how many soldiers can be included per cell as a function of the number of companies on the horizontal axis

and the number of professionals on the vertical axis. Note that, as the number of companies and professionals is only evaluated at integer values, the contour lines are not smooth. It may be clear the number of soldiers per cell decreases as the number of companies and/or professionals increase.

The contour plot at the right shows relative efficiencies of designs with $n_{2A\max} = 30$ and $n_{2B\max} = 30$. The optimal design has relative efficiency equal to 1 and is indicated by a dot: $n_{2A} = 30$ and $n_{2B} = 12$ with $n_1 = 4.2$ and $\text{var}(\hat{\gamma}_1) = .0420$. The noninteger value n_1 may be rounded downward to $n_1 = 4$ such that the total costs are 2,430 and $\text{var}(\hat{\gamma}_1) = .0421$. As the contour lines are rather vertical, any design with $n_{2A} = 29$ or 30 has a high efficiency.

The optimal design may not always be feasible in practice and one may want to use a conditional optimal design by fixing one of the three sample sizes beforehand. Figure 3 shows the conditional optimal designs for various scenarios. The two top graphs relate to the scenario where the number of companies is fixed to $\tilde{n}_{2A} = 20$. The optimal design (rounded to integer values) is found at $n^*_{2B} = 15$ professionals with $n^*_1 = 5$ soldiers per cell. The lines are fairly horizontal, and the relative efficiencies are rather high, which is not surprising given the fairly vertical lines in the right panel of Figure 2.

In the second scenario, the number of professionals is fixed to $\tilde{n}_{2B} = 16$. The two middle graphs show the conditional optimal design (rounded to integer values) is found at $n^*_{2A} = 30$ companies with $n^*_1 = 2$ soldiers per cell. The lines of the relative efficiencies are rather steep, which implies a large loss of efficiency if the chosen design deviates much from the optimal design.

In the third scenario, $\tilde{n}_1 = 10$ soldiers per cell are included in the trial. The two bottom graphs in Figure 3 show the optimal design is found at $n^*_{2A} = 30$ companies and $n^*_{2B} = 5.9$ professionals. Again, a large loss of efficiency is observed if the chosen design deviates much from the optimal design. Rounding to integer values such that the budget is not exceeded gives an optimal design $n^*_{2A} = 30$ companies and $n^*_{2B} = 5$ professionals. This implies each professional treats a total of $n_1 n^*_{2A} = 300$ soldiers. Whether this design is feasible depends on the number of sessions per soldier and the amount of time that is available for the intervention. If the amount of soldiers per professional is too large, one may include fewer units and use the part of the budget that thus becomes available to increase the number of professionals.

Partial Random Cross-Classification

The complete cross-classification of the previous section implicitly requires all health professionals deliver both the intervention and control. This was more or less the case in the study by Mulligan et al. (2012), where one type of health professional delivered the control only and all other types delivered both control

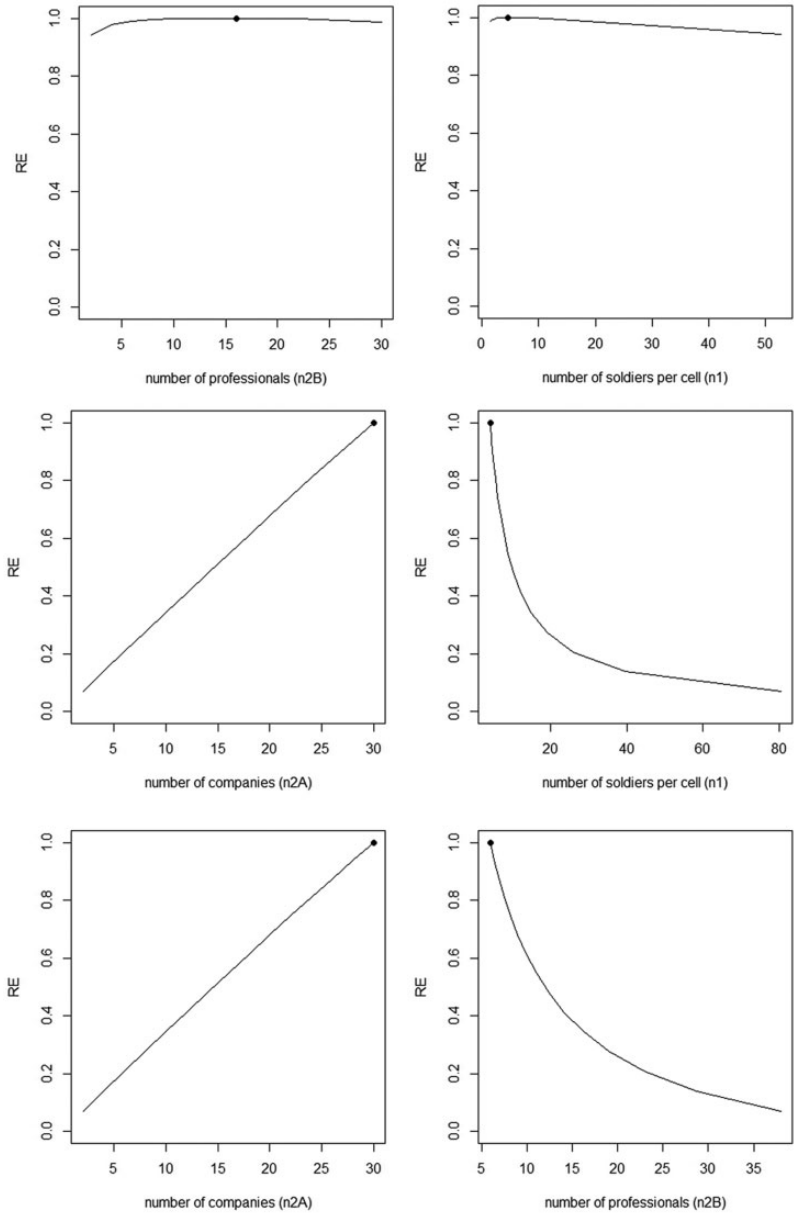

FIGURE 3. *Conditional optimal designs and relative efficiencies of alternative designs. Top graphs:* $\tilde{n}_{2A} = 20$. *Middle graphs:* $\tilde{n}_{2B} = 16$. *Bottom graphs:* $\tilde{n}_1 = 10$. *The dot represents the optimal design.* $C = 2{,}500$, $c_{2A} = 15$, $c_{2B} = 45$, $\rho_u = .3$, $\rho_v = .1$, *and* $\rho_w = .05$.

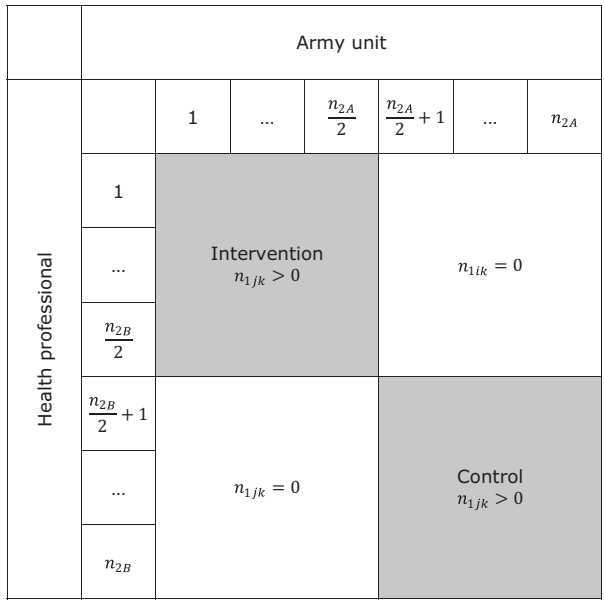

FIGURE 4. *Cross-tabulation of sample size per army unit and health professional in a partial cross-classification.*

and intervention. Such a design may result in contamination of the control group through the health professional and hence an underestimate of the effect of treatment and a reduced power (Moerbeek, 2005). To help ensure intervention fidelity, all intervention sessions were observed by a member of the research team.

Another means to minimize contamination is restricting each professional to offer either the intervention or control but not both. This is a so-called partial random cross-classified design since some of the cells have $n_{1jk} = 0$. Note that we now use subscripts j and k since the number of soldiers per cell is no longer a constant.

For instance, consider a design in which clusters $1, \ldots, n_{2A}/2$ are randomized to the intervention and treated by health professionals $1, \ldots, n_{2B}/2$, and clusters $n_{2A}/2 + 1, \ldots, n_{2A}$ are randomized to the control and treated by health professionals $n_{2B}/2 + 1, \ldots, n_{2B}$. The cross tabulation of n_1 per cell in Figure 4 reveals two blocks for which each cell has $n_{1jk} > 0$, one for the intervention and one for the control. In the other two blocks, each cell has $n_{1jk} = 0$.

In the remainder of this section, we assume a balanced design with $\tilde{n}_1$ soldiers per cell in the first set of two blocks. The variance of the treatment effect estimator of this design is

$$\begin{aligned}\text{var}(\hat{\gamma}_1) &= \frac{4(\sigma_e^2 + \frac{1}{2}\tilde{n}_1 n_{2B}\sigma_u^2 + \frac{1}{2}\tilde{n}_1 n_{2A}\sigma_v^2 + \tilde{n}_1\sigma_w^2)}{\tilde{n}_1 n_{2A} n_{2B}}, \\ &= \frac{4(\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_e^2)(1 + (\frac{1}{2}\tilde{n}_1 n_{2B} - 1)\rho_u + (\frac{1}{2}\tilde{n}_1 n_{2A} - 1)\rho_v + (\tilde{n}_1 - 1)\rho_w)}{\tilde{n}_1 n_{2A} n_{2B}},\end{aligned} \tag{11}$$

which includes the term $\tilde{n}_{1jk} n_{2A}\sigma_v^2$ in the numerator since the intervention condition is a between-professional rather than a within-professional factor. As each health professional delivers only one condition, he or she is crossed by just half of the clusters, and the design is less optimal than the design with a complete random cross-classification.

To illustrate the loss of efficiency, consider a study with $n_{2A} = 30$ army units, $n_{2B} = 30$ health professionals, and $n_1 n_{2B} = 90$ soldiers per army unit (i.e., the total sample size is $n_1 n_{2A} n_{2B} = 2{,}700$). The same intraclass correlations as above are considered: $\rho_u = .3$, $\rho_v = .1$, and $\rho_w = .05$, and the total variance is scaled to 1. For the complete cross-classification, $\text{var}(\hat{\gamma}_1) = .041$; for the partial cross-classification, $\text{var}(\hat{\gamma}_1) = .054$. Taking the ratio of these two variances gives $\text{RE} = .041/.054 = .76$, which implies a considerable loss of efficiency of the partial cross-classified design. The complete cross-classification can also be compared to the nested design, where each army unit has its own health professional. Such a design may have practical advantages, for instance, the management staff of the army unit has to communicate with just one health professional. In this design, the variance components σ_u^2, σ_v^2, and σ_w^2 cannot be estimated separately, and this further decreases the efficiency of the design: $\text{var}(\hat{\gamma}_1) = .061$. The relative efficiency of the nested versus the complete cross-classified design is $\text{RE} = .041/.061 = .67$. In other words, the partial cross-classified design only slightly outperforms the nested design.

Unbalanced Designs

The focus of this section is on random cross-classifications that exist naturally, for instance, the cross-classification of schools and neighborhoods or that of primary and secondary schools. In such cases, the design is likely to be unbalanced, and subsampling to achieve a balanced design is not always possible or justified.

Consider as an example a trial of the long-term effects of a smoking-prevention intervention that was offered to pupils in eighth grade of elementary school in the Netherlands (Ausems, 2003). The last follow-up measurement was taken when all pupils had already transferred to secondary school. It is important to consider the between-secondary school variability while estimating the effect of the intervention on the long-term measurements, since secondary schools may

vary with respect to their policy toward smoking, availability of cigarettes in the school, and its neighborhood and peer pressure among pupils.

The random cross-classification should also be taken into account while designing the study. The R code, available from the first author, can be used to perform a simulation study to calculate the power to detect a treatment effect. This code can handle any number of primary and secondary schools and any number of pupils per primary–secondary school combination. The sample sizes are provided by an external file crossclassdata.txt in which each primary school is represented by a row, each secondary school is represented by a column, and the values in the cells give the number of pupils per primary–secondary school combination. For each simulated data set, random assignment of primary schools to treatment conditions is done, and data are generated and estimated on the basis of equation (1) using the function lmer from the package lme4 (Bates, Mächler, Bolker, & Walker, 2015). The empirical power is calculated as the proportion of data sets for which the null hypothesis is rejected.

To actually use the R code, one must have insight in the distribution of the cell sizes across the primary and secondary schools. For this purpose, one may use administrative data from municipalities or school districts or use data from past research. We use cell sizes from the study by Paterson (1991), where 3,435 pupils are nested in 148 primary schools crossed by 19 secondary schools. In addition, estimates of the intraclass correlation coefficients and the standardized effect size of the primary school intervention must be given. The sizes of these model parameters are given in the first few lines of the R code. We use $\rho_u = \rho_v = \rho_w = .05$; such values are rather common in school-based smoking prevention interventions. The empirical power to detect a small standardized effect size .2 based on 1,000 simulated data set (seed 834) is equal to an acceptable .847.

Conclusion and Discussion

Data in cluster randomized trials do not always have a pure hierarchical structure with subjects nested within clusters. In a random cross-classification structure, the clusters are crossed with clusters of another type. Previous research has shown a data analysis that ignores a random cross-classification data may result in biased results (Gilbert et al., 2016; Luo & Kwok, 2009; Meyers & Beretvas, 2006). It is also important that a random cross-classification is taken into account while designing a cluster randomized trial. This article presents formulas for the sample sizes such that a desired power level is achieved, and optimal sample size equations given a budgetary constraint.

It should be mentioned that this article provides sample size guidelines from a power and optimal design perspective. There are other criteria, guidelines, and recommendations to make decisions with respect to sample size in cluster randomized trials. For instance, Snijders and Bosker (2012) recommend to use at

least 20 clusters if the aim is to generalize inferences to the population of clusters. They mention that with as few as 10 clusters, the data will contain only scant information about the population. Hayes and Moulton (2009) recommend using at least four clusters per treatment condition to ensure a valid analysis. This recommendation is also included in the extension of the Consort 2010 statement to cluster randomized trials (Campbell, Piaggio, Elbourne, & Altman, 2012). Sample sizes should also be chosen such that model parameters and their standard errors are estimated with small bias, and the coverage of confidence intervals is within acceptable limits. Over the past two decades, much simulation research has been done to provide sample size guidelines for frequentist estimation (Maas & Hox, 2005) and Bayesian estimation (Hox, Van De Schoot, & Matthijsse, 2012). These studies focused on perfect hierarchies; future simulation studies should focus on imperfect hierarchies such as random cross-classifications.

The optimal design methodology is also applicable in trials where randomization to treatment conditions is done at the individual level, but treatment is offered by health professionals. Such designs are referred to as individually randomized group treatment trials (Pals et al., 2008). An example is a trial on the effectiveness of surgery for lower back pain, where patients are not only nested within surgeons but also within physiotherapists such that surgeons and therapists are crossed with each other. Patients treated by the same health professional will respond more alike due to therapist effects arising from experience, skills, compassion, adherence to protocols, and so on. Random cross-classifications are further encountered in group treatment trials where several therapist deliver treatment (Roberts & Walwyn, 2013).

This article assumes a constant effect of treatment over therapists, and the multilevel cross-classified was used to take nested and crossed data structures into account. If treatment effects are not constant, however, multilevel models can bias estimated treatment effects, and ordinary least squares with cluster adjusted standard errors might be preferred. Furthermore, we implicitly assumed the subjects to be randomly assigned to therapists. In practice, the assignment of subjects to therapists is often not experimentally controlled, and the multilevel cross-classified model can yield biased standard errors of treatment effect estimates because parameters required to compute the standard errors are not identifiable from the data. So multilevel cross-classified models are used to adjust standard errors in case the data have a nested and cross-classified structure, but they are not always the correct approach to take. Weiss, Lockwood, and McCaffrey (2016) provide suggestions for mitigating the bias due to nonrandom assignment of therapists to subjects.

To apply the optimal design methodology, one must be able to provide prior estimates of the intraclass correlation coefficients of the two crossed random factors. For cluster randomized trials with crossed random effects, such estimates are hardly available in the literature, but at least one can use estimates from

cluster randomized trials without a random cross-classification to get some insight in to the sizes of the intraclass correlation coefficients. A large overview of papers that report estimates of intraclass correlation coefficients is presented by Moerbeek and Teerenstra (2016). For cluster randomized trials without a random cross-classification, it has been shown that optimal designs are rather robust to incorrect prior estimates of the intraclass correlation coefficient (Korendijk, Moerbeek, & Maas, 2010). Furthermore, there are approaches to deal with uncertainty in the intraclass correlation coefficient in the design phase of a trial, such as maximin optimal designs (Van Breukelen & Candel, 2015) and internal pilot designs (Lake, Kammann, Klar, & Betensky, 2002; Van Schie & Moerbeek, 2014).

The balanced design with equal number of subjects per cell is most efficient but cannot always be achieved in practice. The relative efficiency of unbalanced designs has been studied for cluster randomized trials without a crossed random factor (Van Breukelen & Candel, 2012; Van Breukelen et al., 2007). Sampling 10% more clusters often suffices in accounting for unbalancedness. In designs with cross-classified data, the unbalancedness may occur across both random factors and increasing the sample size by 10% may probably not suffice. As Lai and Kwok (2014) already remarked, closed forms of the treatment effect estimator and its variance are very complex functions of the cell sizes. Future research should focus on the loss of efficiency due to an unbalanced design. As for now, the R code, available from the first author, may be used to evaluate the power of unbalanced designs.

Over the past three decades, much attention has been paid to the design and analysis of cluster randomized trials with a pure hierarchical structure. In recent years, a shift in attention has been made to data structures that are not purely hierarchical, such as cross-classified data. Partially cross-classified data were studied in this article but can also be found in studies where some of the subjects are nested in one random factor and others in two cross-classified random factors (Luo, Cappaert, & Ning, 2015). An example is a study on the effect of after-school programs, where pupils in the control are nested in schools and pupils in the intervention are nested within schools crossed by after-school programs. As for now, optimal design recommendations for such trials are lacking. Another example is multiple membership models, see Chung and Beretvas (2012), Luo and Kwok (2012) for the implications of ignoring multiple membership structures, and Roberts and Walwyn (2013) for optimal design methodology. Thus far, the main focus has been on linear models with continuous outcome scores; future research should focus on dichotomous outcome scores.

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.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ali, M., Thiem, V. D., Park, J.-K., Ochiai, R. L., Canh, D. G., Danovaro-Holliday, M. C., . . . Acosta, M. J. (2007). Geographic analysis of vaccine uptake in a cluster-randomized controlled trial in Hue, Vietnam. *Health & Place*, *13*, 577–587. doi:10.1016/j.healthplace.2006.07.004

Ausems, M. (2003). *Smoking Prevention. Comparing in-school, tailored out-of-school, and booster interventions* (Unpublished doctoral dissertation). Maastricht University, the Netherlands.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. F. (2012). Consort 2010 statement: Extension to cluster randomised trials. *British Medical Journal*, *345*, e5661. Retrieved from https://doi.org/10.1136/bmj.e5661.

Campbell, M. J., & Walters, S. J. (2014). *How to design, analyse and report cluster randomised trials in medicine and health related research*. Chichester, England: Wiley.

Castro, C. A., Adler, A. B., McGurk, D., & Bliese, P. D. (2012). Mental health training with soldiers four months after returning from Iraq: Randomization by platoon. *Journal of Traumatic Stress*, *25*, 376–383. doi:10.1002/jts.21721

Chung, H., & Beretvas, S. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65*, 185–200. doi:10.1111/j.2044-8317.2011.02023.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cunningham, T., & Johnson, R. (2012). Design effects for sample size computation in three-level designs. *Statistical Methods in Medical Research*, *25*, 505–519.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, England: Edward Arnold.

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research*. Chichester, England: Wiley.

Gail, M. H., Mark, S. D., Carroll, R. J., & Green, S. B. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, *15*, 1069–1092. doi:10.1002/(SICI)1097-0258(19960615)15:11¡1069::AID-SIM220¿3.0.CO;2-Q

Gilbert, J., Petscher, Y., Compton, D. L., & Schatschneider, C. (2016). Consequences of misspecifying levels of variance in cross-classified longitudinal data structures. *Frontiers in Psychology*, *7*, 695. doi:10.3389/fpsyg.2016.00695

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*, 364–375. doi:10.1177/0049124194022003005

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, England: Wiley.

Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomised trials*. Boca Raton, FL: CRC Press.

Hedges, L., & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics*, *39*, 257–281.

Heo, M., & Leon, A. C. (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*, *64*, 1256–1262.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hox, J., Van De Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87–83.

Konstantopoulos, S. (2009). Incorporating costs in power analysis for three-level cluster randomized designs. *Evaluation Review*, *33*, 335–357. doi:10.1177/0193841X09337991

Korendijk, E. J. H., Moerbeek, M., & Maas, C. J. M. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics*, *35*, 566–585. doi:10.3102/1076998609360774

Lai, M. H. C., & Kwok, O. M. (2014). Standardized mean differences in two-level cross-classified random effects models. *Journal of Educational and Behavioral Statistics*, *39*, 282–302. doi:10.3102/1076998614532950

Lake, S., Kammann, E., Klar, N., & Betensky, R. A. (2002). Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, *21*, 1337–1350. doi:10.1002/sim.1121

Lemme, F., Van Breukelen, G. J. P., & Berger, M. P. F. (2015). Efficient treatment allocation in two-way nested designs. *Statistical Methods in Medical Research*, *24*, 494–512. doi:10.1177/0962280213502145

Luo, W., Cappaert, K. J., & Ning, L. (2015). Modelling partially cross-classified multilevel data. *British Journal of Mathematical & Statistical Psychology*, *68*, 342–362. doi:10.1111/bmsp.12050

Luo, W., & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*, 182–212. doi:10.1080/00273170902794214

Luo, W., & Kwok, O. (2012). The consequences of ignoring individuals' mobility in multilevel growth models—A Monte Carlo study. *Journal of Educational and Behavioral Statistics*, *27*, 31–46. doi:10.3102/1076998610394366

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*, 473–497. doi:10.1207/s15327906mbr4104

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*, 129–149. doi:10.1207/s15327906mbr3901

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable? *The American Statistician*, *59*, 72–78. doi:10.1198/000313005X20727

Moerbeek, M., & Teerenstra, T. (2016). *Power analysis of trials with multilevel data*. Boca Raton, FL: CRC Press.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25*, 271–284. doi:10.2307/1165206

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *The Statistician*, *50*, 1–14. doi:10.1081/STA-200056839

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003a). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*, 341–350. doi:10.1016/S0895-4356(03)00007-6.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003b). Optimal sample sizes in experimental designs with individuals nested within clusters. *Understanding Statistics*, *2*, 151–175. doi:10.1207/S15328031US0203

Mulligan, K., Fear, N. T., Jones, N., Alvarez, H., Hull, L., Naumann, U., . . . Greenberg, N. (2012). Postdeployment battlemind training for the U.K. armed forces: A cluster randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *80*, 331–341.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, *11*, 103–130. doi:10.1076/0924-3453(200003)11:1;1–A;FT103

Pals, S. P., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually randomized group treatment trials: A critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health*, *98*, 1418–1424. doi:10.2105/AJPH.2007.127027

Paterson, L. (1991). Socio economic status and educational attainment: A multidimensional and multilevel study. *Evaluation and Research in Education*, *5*, 97–121.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19*, 337–350. doi:10.3102/10769986019004337

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized studies. *Psychological Methods*, *2*, 173–185. doi:10.1037/1082-989X.2.2.173

Roberts, C., & Walwyn, R. (2013). Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine*, *32*, 81–98. doi:10.1002/sim.5521

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*, 1–15. doi:10.1348/000711008X398968

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.

Teerenstra, S., Moerbeek, M., Van Achterberg, T., Pelzer, B. J., & Borm, G. F. (2008). Sample size calculations for 3-level cluster randomized trials. *Clinical Trials*, *5*, 486–495. doi:10.1177/1740774508096476

Van Breukelen, G. J. P., & Candel, M. J. J. M. (2012). Comments on "Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests." *Statistics in Medicine*, *31*, 397–400. doi:10.1002/sim.4449

Van Breukelen, G. J. P., & Candel, M. J. J. M. (2015). Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Statistical Methods in Medical Research*, *24*, 540–556. doi:10.1177/0962280211421344

Van Breukelen, G. J. P., Candel, M. J. J. M., & Berger, M. P. F. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, *26*, 2589–2603. doi:10.1002/sim.2740

Van Schie, S., & Moerbeek, M. (2014). Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Statistics in Medicine*, *33*, 3253–3268. doi:10.1002/sim.6172

Walwyn, R., & Roberts, C. (2010). Therapist variation within randomised trials of psychotherapy: Implications for precision, internal and external validity. *Statistical Methods in Medical Research*, *19*, 291–315.

Weiss, M. J., Lockwood, J. R., & McCaffrey, D. F. (2016). Estimating the standard error of the impact estimator in individually randomized trials with clustering. *Journal of Research on Educational Effectiveness*, *9*, 421–444.

Authors

MIRJAM MOERBEEK is an associate professor at Utrecht University, PO Box 80140, 3508 TC Utrecht, the Netherlands; email: m.moerbeek@uu.nl. Her research interests are optimal design and statistical power analysis, in particular for multilevel and survival data.

MARYAM SAFARKHANI is a senior statistician at MSD, PO Box 20, 5340 BH Oss, the Netherlands; email: maryam.safarkhani@merk.com. Her research interests are optimal design and statistical power analysis, in particular for longitudinal data.

Manuscript received September 12, 2016
First revision received February 20, 2017
Second revision received May 29, 2017
Accepted August 13, 2017