

RESEARCH ARTICLE

Maximin design of cluster randomized trials with heterogeneous costs and variances

Gerard J. P. van Breukelen  | Math J. J. M. Candel 

Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands

Correspondence

Gerard J. P. van Breukelen, Department of Methodology and Statistics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands.

Email: gerard.vbreukelen@maastrichtuniversity.nl



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Cluster randomized trials evaluate the effect of a treatment on persons nested within clusters, with clusters being randomly assigned to treatment. The optimal sample size at the cluster and person level depends on the study cost per cluster and per person, and the outcome variance at the cluster and the person level. The variances are unknown in the design stage and can differ between treatment arms. As a solution, this paper presents a Maximin design that maximizes the minimum relative efficiency (relative to the optimal design) over the variance parameter space, for trials with two treatment arms and a quantitative outcome. This maximin relative efficiency design (MMRED) is compared with a published Maximin design which maximizes the minimum efficiency (MMED). Both designs are also compared with the optimal designs for homogeneous costs and variances (balanced design) and heterogeneous costs and homogeneous variances (cost-conscious design), for a range of variances based upon three published trials. Whereas the MMED is balanced under high uncertainty about the treatment-to-control variance ratio, the MMRED then tends towards a balanced budget allocation between arms, leading to an unbalanced sample size allocation if costs are heterogeneous, similar to the cost-conscious design. Further, the MMRED corresponds to an optimal design for an intraclass correlation (ICC) in the lower half of the assumed ICC range (optimistic), whereas the MMED is the optimal design for the maximum ICC within the ICC range (pessimistic). Attention is given to the effect of the Welch–Satterthwaite degrees of freedom for treatment effect testing on the design efficiencies.

KEYWORDS

cluster randomized trials, cost function, heterogeneous variance, maximin design, optimal design

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

1 | INTRODUCTION

The effects of a new treatment or intervention in medical, health, and educational research are preferably assessed with a randomized experiment or randomized trial (RCT) in which individuals (patients, students) are randomly assigned to treatment A or B, or treatment or control, and both groups are compared on an outcome of interest after treatment. But individual randomization is sometimes infeasible. To compare different teaching methods or lifestyle interventions in the classroom, we can randomize schools, or classes, but rarely students. Individual randomization can also be undesirable. To compare two types of psychotherapy for depression, we may need to randomize therapists instead of patients to prevent treatment contamination that may arise if a therapist has to switch between two treatments. A similar objection to individual randomization may apply in studies comparing two methods of patient counseling on diabetes or COPD in family practice. If individual randomization is impossible or undesirable, cluster randomization may be the best option. In a cluster-randomized trial (CRT), organizational units (e.g., schools or therapists) are randomly assigned and all individuals in the same unit (e.g., students or patients) are given the same treatment (Donner & Klar, 2000; Hayes & Moulton, 2009; Murray, 1998). CRTs are encountered in family medicine, health promotion, and mental health, among others.

The price of cluster randomization is a lower power and precision compared with individual randomization. This is because outcome variation between clusters in the same treatment arm leads to intraclass correlation (ICC, a correlation between observations in the same cluster), which increases the sampling variance of the treatment effect estimator. This so-called design effect (DE) can be as large as two or three, even if outcome variation between clusters is small relative to that between individuals within clusters (Van Breukelen & Candel, 2018). This makes the optimal design of CRTs important. Here, optimal design means choosing that combination of sample size at each level (number of clusters, number of persons per cluster) that minimizes the sampling variance of the treatment effect estimator and thus maximizes precision and power, for a given study budget. The optimal sample size for a CRT with a quantitative outcome has been presented by Raudenbush (1997) and Moerbeek et al. (2000), among others. Published optimal sample size equations assume that the outcome variance at each design level (cluster, individual) and thereby also the ICC is known, and the same in both treatment arms, and that the study cost per cluster, respectively, per individual is the same in both arms as well. These assumptions are problematic for several reasons. First of all, trials are run to test hypotheses about the *unknown* mean outcome difference between treatments, making the assumption of known variances unrealistic. Further, a treatment which affects the mean of an outcome variable can also be expected to affect its variance, making the assumption of homogeneity of variance realistic only under the null hypothesis. Published evidence on this assumption is scarce as homogeneity is routinely assumed in data analysis and rarely tested and reported explicitly, but there is some evidence for heterogeneity in CRTs (for some examples, see Roberts & Roberts, 2005; ; Cheyne et al., 2008; Adachi et al., 2013; Santos et al., 2020), and even more so in clinical psychology (Grissom, 2000). Finally, with respect to costs, the study costs for sampling, treating, and measuring clusters and persons in one treatment arm may be different from those in the other arm.

The assumption of a known ICC is made in optimal design because that design depends on the ICC value and changing the ICC value changes the optimal design (i.e., it is a locally optimal design [LOD]). The assumption has been relaxed in three different ways, respectively, group sequential design (Lake et al., 2002; Van Schie & Moerbeek, 2014), Bayesian design (Rotondi & Donner, 2009), and Maximin design (Van Breukelen & Candel, 2015). The assumption of homogeneity of costs between the treatment arms of a CRT was relaxed by Liu (2003), and the assumption of homogeneity of variances was dropped in Lemme et al. (2016). Both Lemme et al. and Liu assumed the variances to be known in the design stage, however. This is a problem because the statistical analysis of a CRT with two treatment arms already involves four variance parameters (one per arm at the cluster level, one per arm at the person level) and the optimal design is a function of those four parameters. Allowing the variances to be heterogeneous and unknown, Candel and Van Breukelen (2015) therefore derived Maximin designs for CRTs given a fixed sample size per cluster which was allowed to differ between treatment arms, and Van Breukelen and Candel (2018) generalized this into Maximin designs that optimized the sample size per cluster. Wu et al. (2017) derived Maximin designs for CRTs with a binary outcome, but they assumed a fixed instead of optimal sample size per cluster which was the same for both treatment arms.

Now, Wu et al. used a Maximin relative efficiency (RE) criterion, whereas Candel and Van Breukelen (2015) and Van Breukelen and Candel (2018) used a Maximin efficiency criterion. Maximin efficiency design (MMED) maximizes the efficiency (i.e., minimizes the sampling variance) of the treatment effect estimator in the worst-case scenario, that is, for those true unknown variance parameter values that give the minimum efficiency (maximum sampling variance), hence the name Maximin (or minimax). This has the advantage of guaranteeing a prespecified level of power and precision for treatment effect evaluation across a prespecified plausible variance parameter space. Maximin relative efficiency design

(MMRED) maximizes not the minimum efficiency, but the minimum RE, across the plausible variance parameter space. Here, relative means relative to the optimal design for a given point in the parameter space. This criterion, also known as Minimax Regret, gives a design that stays close to the optimal design across the plausible variance parameter space. These two Maximin criteria are both encountered in optimal design literature (see, e.g., Berger & Wong, 2009, p. 104, 119, 249, 292; Dette et al., 2006; King & Wong, 2000; Muller, 1995; Pronzato & Walter, 1988; Sitter, 1992; Wiens, 2019), but give different results. For the case of a CRT with a quantitative outcome, homogeneous variance, and unknown ICC this was shown in Van Breukelen and Candel (2015), with the MMED being the optimal design for the largest possible ICC, and the MMRED the optimal design for a certain ICC value in the lower half of the ICC range.

The purpose of this paper is threefold. First, to derive the MMRED for a CRT under the same conditions as in Van Breukelen and Candel (2018) for the MMED criterion, that is, a two-arm CRT with a quantitative outcome, heterogeneous known costs, and heterogeneous *unknown* variances, optimizing both the sample size per cluster and the number of clusters. Second, to compare this design with the MMED. Third and last, to compare it with the design obtained by assuming homogeneity of variances and costs (balanced design), and with the design obtained by assuming homogeneity of variance but not of costs (cost-conscious [cc] design). The reason for considering homogeneous variances and heterogeneous costs is that study costs, unlike outcome variances, can be known in the design stage, thereby allowing to take cost heterogeneity into account in the design stage in a simple way. If costs are unknown and assumed to be homogeneous for that reason, then all designs in this paper will be seen to be balanced, at least assuming the same ICC range for both treatment arms. The case of unknown yet heterogeneous costs is beyond the scope of this paper.

The outline of this paper is as follows. First, some results of three published trials are summarized to give an impression of realistic amounts of variance heterogeneity. Second, the mixed model for analyzing a CRT with a quantitative outcome is specified and the optimal sample size per treatment arm and per design level (number of clusters, number of persons per cluster) is given as a function of the costs and variances per arm per level, subject to a fixed total study budget. Next, the MMED of Van Breukelen and Candel (2018) is summarized, specifically, how it divides the study budget between treatment arms and between clusters and individuals. Subsequently, the MMRED is derived. The MMRED is then compared with the MMED in terms of budget split and sample sizes, as well as with the balanced and costs-conscious designs, for a realistic range of treatment-to-control variances and cost ratios. Further, since heterogeneity of variance and unbalanced treatment allocation affect the degrees of freedom (df) for the test statistic for the treatment effect (Satterthwaite, 1941; Welch, 1938), the effect of these df on the relative efficiencies of the different designs in terms of test power and confidence interval width is evaluated. Throughout the paper, an equal sample size per cluster is assumed for all clusters within the same treatment arm, but not between arms. Cluster size variation within treatment arms can be adjusted for in the design stage in a simple and efficient way. For details, see Van Breukelen and Candel (2012). Finally, to assist the reader in keeping track of all mathematical symbols in this paper, Appendix A lists all symbols, their meaning, and the section where they are first used. Further, all figures and tables in this paper have been produced and can be reproduced, with SPSS code and with R code, which is available as a supplement.

2 | EXAMPLES OF HETEROGENEITY IN PUBLISHED TRIALS

Following are summaries of three CRTs reporting the variance of one or more quantitative outcomes per treatment arm to give an impression of realistic amounts of heterogeneity.

Cheyne et al. (2008) compared a new algorithm for diagnosis by midwives of active labor in primiparous women with usual care with respect to oxytocin use and various other outcomes in a CRT of 14 maternity units and 2320 primiparous women in the United Kingdom. The background of this study was the fact that, of all admissions to labor wards, 30–45% concerned women not yet in labor, and that medical interventions were more often given to these women than to women in labor. Of the study outcomes, three were quantitative and two showed substantial heterogeneity of variance at post-test, with a treated to control standard deviation (SD) ratio of 2.09 for time from admission to delivery, and control to treated SD ratio of 1.56 for time from first admission assessment to delivery (table 6 in Cheyne et al.).

Adachi et al. (2013) performed a CRT with 20 general practitioners and 193 type II diabetes patients in Japan to compare a lifestyle education program with usual care in terms of 20 quantitative outcomes. The program aimed at reducing energy intake at dinners and increasing vegetable consumption at breakfast and lunch, thereby improving self-management of glycemic control. The SD ratio exceeded 1.3 for six outcomes, and exceeded 1.5 for three of these: carbohydrate, protein, and fat intake as proportions of total daily energy intake (control to treated SD ratios 1.88, 1.75, 1.55, see table 2 in Adachi et al.).

Santos et al. (2020) reported a CRT on a physical exercise program aiming at improving control of fatigue among industrial workers in a dairy plant in Brazil, in which 13 sectors (the clusters) and 204 workers participated. Both treatment arms received lifestyle education. The intervention group furthermore participated in progressive resistance exercises, whereas the control group performed the usual physical exercises. Pain intensity, one of the outcomes measured after 4 months, showed a control to treated SD ratio of 1.86 (see table 4 in Santos et al.).

The significance of these heterogeneities was tested as follows. First, since none of these publications split the total outcome variance between cluster-level and person-level variance, and none reported ICCs, significance testing was based on the assumption that the ICC was homogeneous so that the SD ratio based on total variances was also the SD ratio based on person-level variances. Second, for a large sample size per cluster, the sampling variance of the person-level variance estimator only depends on the person-level variance and the sample size (Van Breukelen et al., 2008, section 2.2). The SD ratio could therefore be tested with Bartlett's homogeneity of variance test (the Levene test requires the availability of all raw data). This gave a χ^2_1 statistic between 17 and 36 for all three SD ratios in Adachi and the SD ratio in Santos, and above 150 for the two SD ratios in Cheyne (all $p < 0.0001$).

The next section introduces the mixed model for analysis of a two-arm CRT with heterogeneous variance and the optimal design for such a trial. Subsequently, the problem of local optimality (i.e., dependence of the optimal design on variance parameters that are unknown in the design stage of the trial) is addressed by first summarizing the published MMED and then presenting the new MMRED. After that, the MMRED is compared with the popular balanced design, the cc design, and the MMED, in terms of RE and sample size. In these comparisons, we get back to the three examples above, which all used a balanced design.

3 | OPTIMAL DESIGN UNDER HETEROGENEITY OF COSTS AND VARIANCES

To estimate and test the treatment effect on a quantitative outcome Y in a CRT, the following mixed model can be used:

$$Y_{ij} = \beta_0 + \beta_1 X_j + u_j + e_{ij}. \quad (1)$$

Here, Y_{ij} is the outcome for person i in cluster j , and X_j is the treatment assigned to cluster j (1 = treatment, 0 = control). Parameters β_0 and β_1 are fixed effects, and u_j is a random cluster effect, normally distributed with mean zero and variance σ_u^2 , and e_{ij} is a residual term reflecting person and measurement error effects, normally distributed with mean zero and variance σ_e^2 . The u_j s of different clusters are assumed to be uncorrelated, and the e_{ij} s of different persons, whether within the same cluster or not, are also assumed to be uncorrelated. The ICC is now defined as

$$\rho = \frac{\sigma_u^2}{\sigma_y^2} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}, \quad (2)$$

which is the correlation between the outcomes of any two persons in the same cluster due to the shared cluster effect. Generalizing to the case of heterogeneous variance at each design level (cluster, person) gives four variances, denoted as $\sigma_{u(t)}^2$ and $\sigma_{u(c)}^2$ for the cluster-level variance in the treated and control arm, and as $\sigma_{e(t)}^2$ and $\sigma_{e(c)}^2$ for the person-level variance in each arm. Equivalently, we have two total variances $\sigma_{y(t)}^2$ and $\sigma_{y(c)}^2$ and two ICCs ρ_t and ρ_c .

The aim of a CRT is to estimate the treatment effect β_1 as precisely as possible and to have a maximum power for testing this effect. This requires minimization of the sampling variance of the ML estimator $\hat{\beta}_1$, which is (Van Breukelen & Candel, 2018)

$$\text{Var}(\hat{\beta}_1) = [(n_t - 1)\rho_t + 1] \frac{\sigma_{y(t)}^2}{n_t K_t} + [(n_c - 1)\rho_c + 1] \frac{\sigma_{y(c)}^2}{n_c K_c}. \quad (3)$$

Here, n_t is the sample size per cluster and K_t is the number of clusters, in the treated arm, and n_c and K_c are likewise defined for the control arm. The case of homogeneous variances and homogeneous (balanced) sample sizes is obtained by letting $\sigma_{y(t)}^2 = \sigma_{y(c)}^2$, $\rho_t = \rho_c$, $n_t = n_c$, and $K_t = K_c$. The term $[(n - 1)\rho + 1]$ in each arm in Equation (3) indicates the factor by which the sampling variance of the outcome mean in that arm is inflated by the clustering as expressed by the ICC and is known as the design effect (DE). If there is no clustering effect, that is, if $\rho_t = \rho_c = 0$, then Equation (3) reduces

to the sampling variance of the treatment effect estimator in a classical RCT with individual randomization. In practice, the ICC is usually between 0.01 and 0.10 in health care research (Adams et al., 2004; Eldridge et al., 2004), or up to 0.25 in educational research (Hedges & Hedberg, 2007).

The optimal design is the vector (n_t, K_t, n_c, K_c) which minimizes $\text{Var}(\hat{\beta}_1)$ under the constraint of a fixed total study budget B for treating, sampling, and measuring clusters and persons, and a cost vector (c_t, s_t, c_c, s_c) , where c denotes cost per cluster and s is cost per person (subject). For a given budget split into a budget $B_t = fB$ for the treated arm and a budget $B_c = (1 - f)B$ for the control arm, where $f \in (0, 1)$ is the fraction spent on the treated arm, the optimal design is (Van Breukelen & Candel, 2018)

$$n_t^* = \sqrt{\left(\frac{1 - \rho_t}{\rho_t}\right) \left(\frac{c_t}{s_t}\right)}, K_t^* = \frac{B_t}{(c_t + s_t n_t^*)} \quad (4)$$

for the treated arm, and analogously for the control arm. So the optimal sample size per cluster depends only on the cost ratio c/s and ICC for that arm, and the optimal number of clusters depends on that optimal sample size per cluster, and on the costs and the budget for that arm. Inserting these results into Equation (3) gives after rewriting:

$$\text{Var}(\hat{\beta}_1) = \frac{g_t(\rho_t) \sigma_{y(t)}^2}{B_t} + \frac{g_c(\rho_c) \sigma_{y(c)}^2}{B_c}, \quad (5)$$

where

$$g_t(\rho_t) = \left(\sqrt{\rho_t c_t} + \sqrt{(1 - \rho_t) s_t}\right)^2 \in [\text{Min}(s_t, c_t), \text{Sum}(s_t, c_t)], \quad (6)$$

and analogously for $g_c(\rho_c)$ in the control arm, with the maximum of (6) attained at $\rho_t = c_t/(c_t + s_t)$ and likewise for the control arm. Substituting in Equation (5) that $B_t = fB$ and $B_c = (1 - f)B$, and minimizing then (5) as a function of the fraction $f \in (0, 1)$ gives as the optimal budget split between arms:

$$\frac{f^*}{1 - f^*} = \frac{\sigma_{y(t)} \sqrt{g_t(\rho_t)}}{\sigma_{y(c)} \sqrt{g_c(\rho_c)}}. \quad (7)$$

Combining (5) and (7) results in the following minimum variance of the treatment effect under heterogeneity:

$$\text{Var}^*(\hat{\beta}_1) = \left(\sigma_{y(t)} \sqrt{g_t(\rho_t)} + \sigma_{y(c)} \sqrt{g_c(\rho_c)}\right)^2 / B. \quad (8)$$

4 | MAXIMIN EFFICIENCY DESIGN

Equations (4)–(8) show that the optimal budget split between the two groups (treated, control) and the optimal sample size (number of clusters, number of persons) depend on four cost parameters and four variance parameters. Now, the cost parameters can be known in the design stage, but the variances cannot. This makes optimal design vulnerable to misspecification of the variances. One way to obtain robustness against misspecification is MMED, which consists of the following steps:

1. Specify the parameter space, that is, the region of all plausible values for the unknown vector $(\sigma_{y(t)}^2, \rho_t, \sigma_{y(c)}^2, \rho_c)$ on which the optimal design depends, subject to the following constraints: $(\sigma_{y(t)}^2 + \sigma_{y(c)}^2) \in [V_{\min}, V_{\max}]$, $\rho_t, \rho_c \in [\rho_{\min}, \rho_{\max}]$, and $\sigma_{y(t)}/\sigma_{y(c)} \in [1/u, u]$. Here, the lower and upper bounds, V_{\min} and V_{\max} , and ρ_{\min} and ρ_{\max} , are to be specified by the user, based on prior knowledge. Further, the range $[1/u, u]$ is chosen based on the amount of uncertainty about the heterogeneity of the variance, with $u = 1$ giving homogeneity and large u allowing for much uncertainty.

2. Specify the design space, that is, the set of all candidate sample sizes (n_t, K_t, n_c, K_c) , subject to the budget constraint $K_t(c_t + n_t s_t) + K_c(c_c + n_c s_c) = B$, where the total budget B is to be specified by the user, and subject to the sample size constraint $n_t, K_t, n_c, K_c \geq 1$.
3. For each design in the design space, find its minimum efficiency, or maximum $\text{Var}(\hat{\beta}_1)$, within the parameter space.
4. Select the design with the smallest maximum $\text{Var}(\hat{\beta}_1)$ (Minimax design), or equivalently, the design with the largest minimum efficiency (Maximin design). This is the MMED, which is robust against misspecification of the unknown parameters by optimizing the worst case.

The variance and budget constraints in steps 1 and 2 are needed because, as Equation (8) shows, $\text{Var}(\hat{\beta}_1)$ increases with the variances, and decreases as the budget increases, and infinitely large variances would require an infinitely large budget. Concerning step 1 above, based on published reviews of ICC values (Adams et al., 2004; Eldridge et al., 2004; Hedges & Hedberg, 2007), it can be safely assumed that $\rho_{\max} \leq 0.50$. Concerning step 3 above, combining $\rho_{\max} \leq 0.50$ with the reasonable assumptions $c_t \geq s_t$ and $c_c \geq s_c$, it can be seen that $\rho_t, \rho_c = \rho_{\max}$ is a necessary condition to obtain the worst-case scenario of a maximum $\text{Var}(\hat{\beta}_1)$ by Equations (6) and (8). Another necessary condition is $\sigma_{y(t)}^2 + \sigma_{y(c)}^2 = V_{\max}$. The worst case is now obtained by filling in these two conditions in Equation (5), and then maximizing the resulting $\text{Var}(\hat{\beta}_1)$ as a function of the SD ratio $\sigma_{y(t)}/\sigma_{y(c)}$ within the constraint $\sigma_{y(t)}/\sigma_{y(c)} \in [1/u, u]$. Concerning step 4 of the Maximin procedure, the MMED is obtained by taking the maximum $\text{Var}(\hat{\beta}_1)$ from step 3 and then finding that budget split between arms that minimizes it. This Maximin split is as follows (for a proof, see van Breukelen & Candel, 2018; Appendix B):

$$\frac{f^m}{1-f^m} = p^2 \text{ if } p \in \left[\frac{1}{u}, u \right] \quad (9a)$$

$$\frac{f^m}{1-f^m} = pu \text{ if } p > u \quad (9b)$$

$$\frac{f^m}{1-f^m} = \frac{p}{u} \text{ if } p < \frac{1}{u}, \quad (9c)$$

where $p = \sqrt{g_t(\rho_{\max})/g_c(\rho_{\max})}$. The implication of this budget split is best seen if $c_t/s_t = c_c/s_c$, which gives $p = \sqrt{c_t/c_c} = \sqrt{s_t/s_c}$, the square root of the treatment-to-control cost ratio. If $p \in [1/u, u]$, so the costs are less heterogeneous than the variances can be, then the treatment-to-control budget ratio equals the cost ratio, p^2 , and the design is balanced: $n_t = n_c$ and $K_t = K_c$ (see Equation 4 with $c_t/s_t = c_c/s_c$ and $\rho_t, \rho_c = \rho_{\max}$. If $p \notin [1/u, u]$, so the costs are more heterogeneous than the variances can be, the budget ratio is in-between the cost ratio p^2 and its square root p . In that case, although more budget is allocated to the more expensive arm, the number of clusters is larger in the cheaper arm. So, if $p > u$, we have $B_t > B_c$, yet $K_t < K_c$ (see Equations 9b and 4). Likewise, if $p < 1/u$, we have $B_t < B_c$, yet $K_t > K_c$. The MMED is finally obtained with Equation (4) by using $\rho_t, \rho_c = \rho_{\max}$ and the budget split in Equation (9).

5 | MAXIMIN RELATIVE EFFICIENCY DESIGN

Maximin design based on the efficiency criterion, or equivalently, on the $\text{Var}(\hat{\beta}_1)$ criterion, is safe in considering the worst-case scenario of a maximum $\text{Var}(\hat{\beta}_1)$. However, this scenario may be unlikely to occur, as it requires the ICC to be on the upper boundary of its range, that is, $\rho_t, \rho_c = \rho_{\max}$, and, unless $p \in [1/u, u]$, it also requires the SD ratio to be on a boundary of its range, specifically: $\sigma_{y(t)}/\sigma_{y(c)} = u$ if $p > u$, or $\sigma_{y(t)}/\sigma_{y(c)} = 1/u$ if $p < 1/u$ (for details and proofs, see Van Breukelen & Candel, 2018; Appendix B). In view of this, an alternative is to maximize the RE across the parameter space, here across the ICC range $[\rho_{\min}, \rho_{\max}]$, the SD ratio range $[1/u, u]$, and the range $[V_{\min}, V_{\max}]$ for the variance sum $\sigma_{y(t)}^2 + \sigma_{y(c)}^2$. This leads to the MMRED. Instead of maximizing the minimum efficiency (minimizing the maximum $\text{Var}(\hat{\beta}_1)$), it maximizes the minimum RE, where relative means: as compared to the LOD for a given point in the parameter space, that is, for a given set of variance parameter values $(\sigma_{y(t)}^2, \rho_t, \sigma_{y(c)}^2, \rho_c)$. MMRED thus differs from MMED in its last two steps:

Step 3: For each design D in the design space, derive its minimum RE compared with the LOD over the parameter space, that is, minimize $\text{Var}(\hat{\beta}_1|LOD)/\text{Var}(\hat{\beta}_1|D)$ as a function of the unknown variance parameter values on which the RE depends;

Step 4: Now select that design D which has the highest minimum RE. This is the MMRED, which is robust against misspecification of the variance parameters in the sense that it stays as close to the LOD as possible across the parameter space.

For this RE criterion, the variance sum $\sigma_{y(t)}^2 + \sigma_{y(c)}^2$ is irrelevant (for details, see Appendix B), and so the parameter space is restricted to the Cartesian product of the ICC range $[\rho_{\min}, \rho_{\max}]$ and the SD ratio range $[1/u, u]$. Now, for the case of homogeneous variances ($u = 1$), Van Breukelen and Candel (2015) derived the MMRED sample size per cluster, from which the number of clusters then follows due to the budget constraint. Applying their result per arm in the heterogeneous case gives as MMRED sample size per cluster in the treated arm:

$$n_t^r = \frac{(1 - \rho_{\min})g_t(\rho_{\max}) - (1 - \rho_{\max})g_t(\rho_{\min})}{\rho_{\max}g_t(\rho_{\min}) - \rho_{\min}g_t(\rho_{\max})}, \quad (10)$$

and analogously for the sample size n_c^r in the control arm.

Here, the superscript r indicates the MMRED sample size per cluster as opposed to the MMED (superscript m) or locally optimal sample size (superscript $*$). Equation (10) requires $\rho_{\min} < \rho_{\max}$.

If $\rho_{\min} = \rho_{\max}$, then the ICC is known and the sample size per cluster follows from Equation (4).

To give an impression, $\rho \in [0, 1]$ gives $n^r = c/s$. As Equation (4) shows, this is the LOD for $\rho = s/(c + s)$, which is in the lower half of the ICC range if $c > s$. Taking the more realistic ICC range $\rho \in [0, 0.5]$ gives $n^r = (c + 2\sqrt{cs})/s$, which is the LOD for an ICC below 0.10 if $c > s$, as may be seen by letting $n^r = n^*$, and solving for ρ in Equation (4). In short, the MMRED is the LOD for an ICC value in the lower half of its range, in contrast with the MMED which is the LOD for the maximum ICC.

To derive now the number of clusters per treatment arm according to the MMRED, we need to decide on the budget split between both arms, since the number of clusters is proportional to the budget (see Equation 4). Therefore, let us first write the sampling variance of the treatment effect, $\text{Var}(\hat{\beta}_1)$, as a function of the MMRED sample size per cluster given by Equation (10), and budget split $B_t = fB$, $B_c = (1 - f)B$, with $f \in (0, 1)$:

$$\text{Var}(\hat{\beta}_1) = \frac{h_t(\rho_t)\sigma_{y(t)}^2}{fB} + \frac{h_c(\rho_c)\sigma_{y(c)}^2}{(1-f)B}, \quad (11)$$

where

$$h_t(\rho_t) = [(n_t^r - 1)\rho_t + 1] \left(\frac{c_t + n_t^r s_t}{n_t^r} \right), \quad (12)$$

in the treated arm, and analogously for $h_c(\rho_c)$ in the control arm. Note that the first factor in (12) is the DE (Section 3) and the second factor is the total cost per sampled person including cluster costs. Equation (11) is a rewriting of Equation (3) and has the same structure as Equation (5).

Defining now

$$p_1 = \frac{\sqrt{h_t(\rho_{\min})}}{\sqrt{h_c(\rho_{\max})}}, p_2 = \frac{\sqrt{h_t(\rho_{\max})}}{\sqrt{h_c(\rho_{\min})}}, \quad (13)$$

which are the minimum, respectively, maximum, of $\sqrt{h_t(\rho_t)/h_c(\rho_c)}$ the following MMRED budget ratio can be derived (for a detailed proof, see Appendix B):

$$\frac{f^r}{1 - f^r} = \frac{2p_1 p_2 + \left(\frac{p_1}{u}\right) + p_2 u}{2 + \left(\frac{p_1}{u}\right) + p_2 u}. \quad (14)$$

To show how the MMRED behaves as a function of cost and variance heterogeneity, we focus on a special case where the MMRED is a function of two parameters only, one for cost heterogeneity and one for variance heterogeneity, with both parameters applying at the cluster level as well as at the person level. This case is based on two constraints. First,

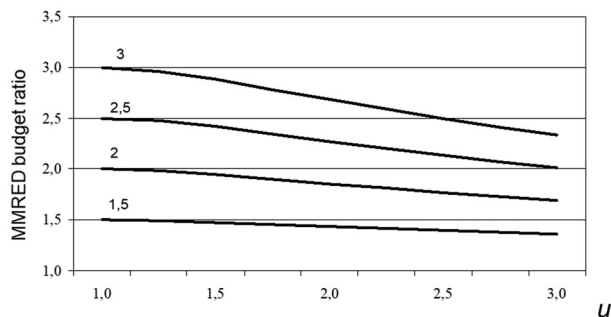


FIGURE 1 Maximin (MMRED) budget ratio treatment-to-control, based on minimum RE criterion, as a function of the SD ratio range $[u^{-1}, u]$ for various p (square root of the treatment-to-control cost ratio)

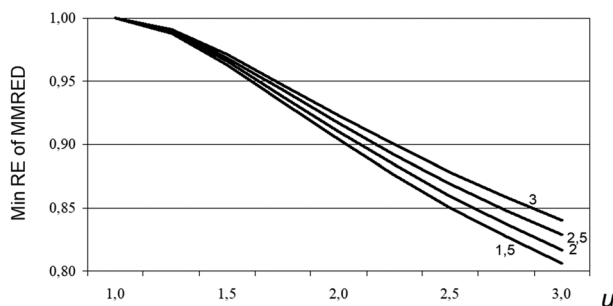


FIGURE 2 Minimum relative efficiency (min RE) of the MMRED, as a function of the range for the unknown SD ratio, $[u^{-1}, u]$, for various p (square root of the treatment-to-control cost ratio)

$c_t/s_t = c_c/s_c$ (homogeneous cost ratio), which gives $n_t = n_c$ (homogeneous cluster size) by Equations (6) and (10), and $p_1 p_2 = c_t/c_c = s_t/s_c$ (the treatment to control cost ratio) by Equations (12) and (13). Second, $\rho_{\min} = \rho_{\max}$ (known and homogeneous ICC), so that the optimal cluster size is given by Equation (4), and that $p_1 = p_2 = \sqrt{c_t/c_c} = \sqrt{s_t/s_c}$, the square root of the treatment-to-control cost ratio, denoted by p . This makes the MMRED a function of two parameters, p and u , just as in the MMED design in the previous section (note: plots for the case $\rho_{\min} < \rho_{\max}$ and $p_1 < p_2$ showed similar results, but contained too many curves).

Figure 1 plots the MMRED budget ratio against u , for various p , with both parameters running from 1 to 3 (implying a variance and cost ratio up to 9), based upon the trials in Section 2. If $u = 1$ (homogeneous variance) we get $f^r/(1 - f^r) = p$ by Equation (14), remembering that $p_1 = p_2$ if the ICC is known. As u and thus variance heterogeneity increases, the budget ratio moves from p towards 1. So, for homogeneous variances ($u \approx 1$), more budget is spent on the more expensive treatment. For heterogeneous variances (large u), the budget split becomes more balanced, giving an unbalanced design, unless $p = 1$ (homogeneous costs). This is different from the MMED budget ratio in Equation (9), which moves not from p to 1, but from p to p^2 , giving a balanced design, as u increases.

Figure 2 plots the minimum RE of the MMRED, which is at least 0.90 for u up to 2 (i.e., a variance ratio between 0.25 and 4), and still at least 0.80 for u up to 3 (variance ratio between 0.11 and 9). Further, the minimum RE increases as the costs become more heterogeneous, but this effect is much smaller than that of heterogeneous variances.

The present results are based on a two-sided interval for the SD ratio. The end of Appendix B points out how the MMRED budget split and its minimum RE change if a one-sided interval is assumed. Since that requires the researcher to specify in advance whether the variance will be larger in the treated arm or the control arm, this topic is not elaborated here.

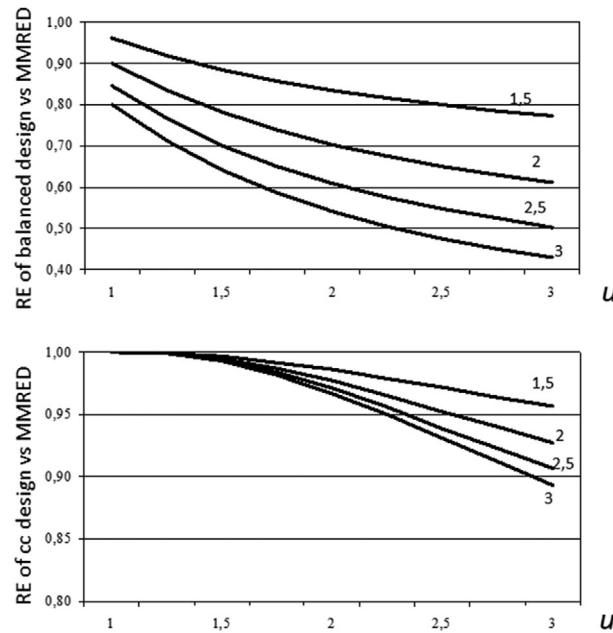


FIGURE 3 RE of the balanced design (upper panel) and the cc design (lower panel) versus the MMRED with respect to the minimum RE criterion as a function of the SD ratio range $[u^{-1}, u]$ and p

6 | EFFICIENCY COMPARISON WITH THE BALANCED DESIGN AND THE COST-CONSCIOUS DESIGN

The MMRED will be compared with the popular balanced design in terms of its RE (this section) and its sample size (next section) for a given budget, for realistic amounts of variance heterogeneity based on the examples in Section 2. Unfortunately, the study cost per cluster and per person is never reported in the authors' experience. Therefore, the same range will be assumed for the cost ratio as for the variance ratio.

To see how much can be gained by using the MMRED, take the popular balanced design, which was used in all three trials summarized in the previous section. For simplicity, the same assumptions are made as in Section 4 for Figures 1 and 2, that is, $c_t/s_t = c_c/s_c$ (homogeneous cost ratio) and $\rho_{\min} = \rho_{\max}$ (known homogeneous ICC), so that $p_1 = p_2 = \sqrt{c_t/c_c} = \sqrt{s_t/s_c}$, the square root of the treatment-to-control cost ratio, denoted as p , and that $n_t = n_c$ with the same value of n in each design (see Equation 4). This makes the comparison between designs dependent on just two parameters: the maximum possible SD ratio u and the square root cost ratio p . Remember, however, that the optimal and Maximin designs in this paper do not require these assumptions (see Equations 4,7,9,10,14). Further, the comparison will allow the SD ratio parameter u to run from 1 to 3 in view of the results in Section 5. Lacking empirical evidence on cost ratios, p will also run from 1 to 3.

From the right half of Equation (4), it follows that the balanced design (where $n_t = n_c$ and $K_t = K_c$) needs as budget split $f^b/(1 - f^b) = p^2$. Inserting this into Equation (B3) of Appendix B gives after rewriting the minimum RE of the balanced design (for the equations, see Appendix C). Below, the MMRED and balanced design will be compared in terms of this minimum RE. First, however, another candidate design is introduced: the cost-conscious (cc) design of Van Breukelen and Candel (2018), which is the optimal design for homogeneous variances and heterogeneous costs. This is a practical alternative to the balanced design because, unlike variances, costs can be known in the design phase and heterogeneity can thus be easily accounted for. From Equations (7) and (9), it follows that the cc design has budget split $f^c/(1 - f^c) = p$ if $c_t/s_t = c_c/s_c$, and the cluster sizes and numbers of clusters per arm then follow from Equation (4). Inserting this budget split into Equation (B3) of Appendix B gives upon rewriting the minimum RE of the cc design (see Appendix C).

A direct comparison between the MMRED and the balanced and cc designs in terms of their minimum RE is given in Figure 3, which shows (a) the RE of balanced versus MMRED and (b) the RE of cc versus MMRED, using as criterion the minimum RE of each design. So, each panel in Figure 3 shows the ratio of two minimum REs, where each of the two minimum REs is relative to the LOD given by Equations (4) and (7). The REs of the balanced and the cc design compared to the MMRED design both decrease as the costs and variances become more heterogeneous. The balanced design is much

less efficient than the MMRED if costs and variances are very heterogeneous, becoming as low as 0.70 if $p = u = 2$ and 0.40 if $p = u = 3$. In contrast, the cc design is quite efficient, with an RE of 0.98 for $p = u = 2$ and 0.89 for $p = u = 3$ (note the different scales on the vertical axes of the two panels). These results can be understood by looking at Figure 1, and remembering that the comparisons concern the case where $c_t/s_t = c_c/s_c$ and $p_1 = p_2$. Under homogeneity of variance ($u = 1$), the budget allocation ratio for the MMRED is then p (see Equation 14), which is the same as for the cc design, whereas the balanced design has a budget ratio p^2 . As heterogeneity of variance increases ($u \rightarrow \infty$), the budget ratio of the MMRED moves towards 1 and away from those of the cc design and especially the balanced design.

These results differ from those in Van Breukelen and Candel (2018) for the MMED and the minimum efficiency criterion. They found the following results for the minimum efficiency criterion, so defining the RE as the ratio of the minimum efficiencies of the designs that are compared, instead of as the ratio of the minimum REs (relative to the LOD):

- I. the RE of the balanced design compared to the MMED *increased* as *variance* heterogeneity increased, such that the balanced design was the MMED if $p \in [u^{-1}, u]$;
- II. the RE of the cc design compared to the MMED *increased* slightly as *cost* heterogeneity increased, and was never below 0.80;
- III. in terms of the minimum efficiency criterion, the balanced design was *more* efficient than the cc design if $p \in [u^{-2}, u^2]$, and less than the cc design else.

What does this mean for the three CRTs in Section 2? All three trials used a balanced design and reported some SD ratios close to 2 (either treated: control, or control: treated). In terms of the minimum efficiency criterion, the balanced design is quite efficient and better than the cc design unless the cost ratio exceeds the variance ratio. In terms of the minimum RE criterion, however, the balanced design is inefficient for an SD ratio of 2, with a RE of only 0.70 if $p = 2$ or 0.55 if $p = 3$. Only for homogeneous costs ($p = 1$) is the balanced design the MMRED.

7 | SAMPLE SIZE COMPARISON

The preceding sections compared the four designs, balanced, cost-conscious, MMED, and MMRED, in terms of the budget allocation ratio and their relative efficiencies. This section compares them in terms of the sample allocation ratio, that is, the number of clusters per treatment arm, as a function of the cost heterogeneity parameter p and the variance heterogeneity parameter u , for a given study budget and using the same assumptions as in the preceding section (which imply $n_t = n_c$, with the same value of n for all four designs).

Based on typical sample sizes in CRTs according to a published review (Adams et al., 2004), 20 clusters per treatment arm are assumed for the balanced design. Based on Section 5, the SD ratio is allowed to vary from 1 to 3, so $u = 3$. Lacking publications of study costs, p is also allowed to vary from 1 to 3. In varying p from 1 to 3, the average (A) of $c_t + n_t s_t$ (cost per treated cluster including person costs) and $c_c + n_c s_c$ (cost per control cluster including person costs) is kept constant. This is done to keep the budget B needed for the balanced design of 20:20 clusters constant for a fair comparison with other designs (for instance, $B = 2000$ and $A = 50$). The sample size for the cc design follows from its budget allocation ratio p which, combined with a cost ratio p^2 , implies a sample allocation ratio of $1/p$. The sample sizes for the MMED and MMRED are similarly computed from their budget ratios as given by Equations (9) and (14).

Table 1 lists the number of clusters per arm as a function of p from 1 to 3 and u from 1 to 3, for all four designs: balanced, cc, MMED, and MMRED. The following trends can be seen. First of all, under cost homogeneity ($p = 1$) all designs are balanced, and under variance homogeneity ($u = 1$) both Maximin designs reduce to the cc design. Second, the MMED is balanced if the variance heterogeneity u is at least as large as the cost heterogeneity p . Third and last, as u and p increase, the MMRED moves away from the balanced design while staying close to the cc design for p and u up to 3. Table 1 assumes 20 clusters per arm for the balanced design but results for other numbers of clusters are easily inferred. For instance, to compare with a balanced design with 10 clusters per arm, divide all numbers in Table 1 by two so that all designs need the same budget.

Given the omnipresence of balanced designs, these results have again a practical implication. If the study costs per cluster and per person are the same for both treatment arms (i.e., $p = 1$), then all four designs in Table 1 are balanced irrespective of variance heterogeneity. But if the costs are heterogeneous, then both Maximin designs are close to the cc design under homogeneity of variance, and the MMRED is so even under heterogeneity of variance. As Figure 3 shows, the RE of the balanced design compared with the MMRED can be quite low then, for instance, 0.70 or even 0.55 for an SD

TABLE 1 Number of clusters per treatment arm according to three alternatives to the balanced design with 20 clusters in each arm for the same study budget and costs, as a function of the square root treated-to-control cost ratio (p) and the range for the square root treated-to-control variance ratio $[1/u, u]$, assuming a known and homogeneous ICC and homogeneous cluster-to-person cost ratio

| p | u | Cost-conscious | | Maximin efficiency | | Maximin relative efficiency | |
|-----|-----|----------------|-------|--------------------|-------|-----------------------------|-------|
| 1 | 1 | 20 | 20 | 20 | 20 | 20 | 20 |
| | 2 | 20 | 20 | 20 | 20 | 20 | 20 |
| | 3 | 20 | 20 | 20 | 20 | 20 | 20 |
| 2 | 1 | 16.67 | 33.33 | 16.67 | 33.33 | 16.67 | 33.33 |
| | 2 | 16.67 | 33.33 | 20 | 20 | 16.25 | 35 |
| | 3 | 16.67 | 33.33 | 20 | 20 | 15.71 | 37.14 |
| 3 | 1 | 16.67 | 50 | 16.67 | 50 | 16.67 | 50 |
| | 2 | 16.67 | 50 | 19.05 | 28.57 | 16.19 | 54.29 |
| | 3 | 16.67 | 50 | 20 | 20 | 15.56 | 60 |

ratio of 2 as in the three CRTs in Section 2, depending on the cost ratio. For researchers planning CRTs, it is thus important to consider study costs per treatment arm as well as possible heterogeneity of variance instead of automatically choosing a balanced design.

8 | DEGREES OF FREEDOM FOR TESTING AND INTERVAL ESTIMATION

Until now, the efficiency of a design was defined in terms of $\text{Var}(\hat{\beta}_1)$, ignoring the complicating factor that, with unknown variances, the test statistic for the treatment effect has a Student t distribution rather than a standard normal distribution. This follows from the equivalence of mixed (multilevel) regression analysis of individual data following Equation (1) with the unpaired t -test of treatment versus control using clusters as units of analysis and cluster means as outcome (see, e.g., Moerbeek et al., 2003). Under heterogeneity of variance, the degrees of freedom for this t -test obey the following expression (Welch, 1938; Satterthwaite, 1941; Welch, 1938):

$$df = \frac{\left(\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}\right)^2}{\left(\frac{1}{K_1-1}\right)\left(\frac{\sigma_1^2}{K_1}\right)^2 + \left(\frac{1}{K_2-1}\right)\left(\frac{\sigma_2^2}{K_2}\right)^2}, \quad (15)$$

where

$\sigma_1^2 = [(n_t - 1)\rho_t + 1](\sigma_{y(t)}/n_t)$ is the sampling variance of the outcome mean in an arbitrary treated cluster, K_1 is the number of treated clusters, and σ_1^2/K_1 is, therefore, the sampling variance of the outcome mean under treatment, and analogously for σ_2^2, K_2 , and σ_2^2/K_2 under control. Equation (15) implies that df varies between the minimum of $K_1 - 1$ and $K_2 - 1$ (if σ_1^2/K_1 is very small relative to σ_2^2/K_2 or vice versa) and their sum (if $\sigma_1^2 = \sigma_2^2$ and $K_1 = K_2$). In practice, df is estimated by replacing σ_1^2 and σ_2^2 with their estimators.

Equation (15) is relevant to the efficiency of the designs in this paper because the df determines the critical value of the test statistic for a given α , thereby also affecting the test power, here denoted by $(1 - \gamma)$. The higher the df , the smaller the critical value and the higher the power, since the sample size needed for the Student t -test is proportional to the factor $(t_{df,1-\gamma} + t_{df,1-\alpha/2})^2$, where $t_{df,1-\gamma}$ is the $100(1 - \gamma)$ -th percentile of the t -distribution with df degrees of freedom and analogously for $t_{df,1-\alpha/2}$ (Julious, 2010). Similarly, the sample size needed for a specific confidence interval width for the treatment effect is proportional to $(t_{df,1-\alpha/2})^2$. The case $df \rightarrow \infty$ gives the standard normal distribution, but the Student t -distribution is very close to that for $df = 100$ or more. This means that especially the efficiency of designs with a small df according to Equation (15) is a bit overestimated by merely considering $\text{Var}(\hat{\beta}_1)$ in comparing different designs. Taking the df effect into account, the RE of a design D1 versus another design D2 can be defined as (for a proof, see Appendix D):

$$\text{RE}(D1 \text{ versus } D2) = \frac{\text{Var}(\hat{\beta}_1|D2)}{\text{Var}(\hat{\beta}_1|D1)} \times \frac{(t_{df2,1-\gamma} + t_{df2,1-\alpha/2})^2}{(t_{df1,1-\gamma} + t_{df1,1-\alpha/2})^2}, \quad (16)$$

TABLE 2 Ratio of the sample size term $(t_{1-\alpha/2} + t_{1-\gamma})^2$ of left design: right design, as a function of the square root treated-to-control cost ratio (p) and the range for the square root treated-to-control variance ratio $[1/u, u]$, for two extreme cases: Actual SD ratio (SDr) = u (left number) and actual SD ratio = $1/u$ (right number), and for two sample sizes: 40 clusters for the balanced design (top table) and 20 clusters (bottom table). Assumptions: Known and homogeneous ICC, homogeneous cluster-to-person cost ratio, two-tailed testing with $\alpha = 0.05$ and power $(1 - \gamma) = 0.90$. Values in boldface differ more than 5% from 1. Values below 1 are in favor of the left design, values above 1 are in favor of the right design

| p | u | Balanced: MME | | Balanced: MMRE | | Cc: MME | | Cc: MMRE | | |
|---|-----|---------------|-------------|----------------|-------------|-------------|-------------|-----------|-------------|---|
| | | SDr = u | SDr = $1/u$ | SDr = u | SDr = $1/u$ | SDr = u | SDr = $1/u$ | SDr = u | SDr = $1/u$ | |
| For a sample size of 40 clusters for the balanced design (20 per arm) | | | | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 0.99 | 0.99 | 0.99 | 0.99 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 0.96 | 1.03 | 1.03 | 0.97 | 1 | 1 | 1 |
| | 3 | 1 | 1 | 0.96 | 1.05 | 1.03 | 0.96 | 0.99 | 1.01 | 1 |
| 3 | 1 | 0.98 | 0.98 | 0.98 | 0.98 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 0.99 | 1.03 | 0.96 | 1.03 | 1.03 | 0.99 | 0.99 | 1 | 1 |
| | 3 | 1 | 1 | 0.96 | 1.06 | 1.03 | 0.94 | 0.99 | 1 | 1 |
| For a sample size of 20 clusters for the balanced design (10 per arm) | | | | | | | | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 0.97 | 0.97 | 0.97 | 0.97 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 1 | 0.92 | 1.07 | 1.08 | 0.93 | 0.99 | 1 | 1 |
| | 3 | 1 | 1 | 0.91 | 1.11 | 1.07 | 0.91 | 0.98 | 1.01 | 1 |
| 3 | 1 | 0.95 | 0.95 | 0.95 | 0.95 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 0.97 | 1.06 | 0.90 | 1.08 | 1.07 | 0.98 | 0.99 | 1 | 1 |
| | 3 | 1 | 1 | 0.90 | 1.14 | 1.08 | 0.88 | 0.97 | 1 | 1 |

where $df2$ is the df for design D2 and $df1$ is the df for design D1. Equation (16), rather than its first factor alone, gives the ratio of study budgets needed for the two designs to be equally powerful in determining a treatment effect. For instance, if the RE as defined by Equation (16) is 2, then D2 needs twice as large budget as D1 to have the same power. Strictly speaking, the budget for D2 needs to increase a bit less than that, because doubling the budget for D2 not only reduces $\text{Var}(\hat{\beta}_1|D2)$ with 50% but also reduces the term $(t_{df2,1-\gamma} + t_{df2,1-\alpha/2})^2$ by increasing the sample size and df of D2. However, the magnitude of the latter effect depends on the df of D2 before the budget increase and is thus not easily incorporated into Equation (16).

To show the effects of the df differences on the design efficiencies, Table 2 lists the last factor of Equation (16), so the ratio of the factor $(t_{1-\gamma} + t_{1-\alpha/2})^2$ for the balanced and cc designs to the factor for the MMED and MMRED, as a function of the square root cost ratio p and the square root maximum variance ratio u , under the same assumptions as in Sections 6 and 7. The top table assumes a total of 40 clusters for the balanced design (20 per arm), and the bottom table a total of 20 clusters (10 per arm). For each design pair (columns) and each (p,u) pair (rows) there are two ratios: one for the case where the actual SD ratio in the trial is u (maximum heterogeneity, the more expensive arm has the larger variance), and one where the actual SD ratio is $1/u$ (maximum heterogeneity, the more expensive arm has the smaller variance). It follows from Table 2 that, for a budget such that the balanced design has 20 clusters per arm, the ratio deviates less than 5% from 1 for all combinations except the most heterogeneous scenario ($p = u = 3$), where the deviation can be 6%. Note, however, that this requires both the cost ratio p^2 and the maximum variance ratio u^2 to be at least as large as 9, which is quite extreme. In contrast, for a budget such that the balanced design has 10 clusters per arm, deviations of more than 5% frequently occur, and even deviations exceeding 10% occur. This suggests that, when the number of clusters is small, the effect of using a t -test versus z -test on the relative efficiencies of the designs cannot be ignored. Table 2 applies to two-tailed

testing with $\alpha = 5\%$ and a power of 90%. For confidence interval estimation, the same equation can be used by taking a power of 50% so that $t_{1-\gamma} = 0$. This is because the confidence interval width is proportional to $t_{1-\alpha/2} \sqrt{\text{Var}(\hat{\beta}_1)}$. The results for that (not shown) are very similar to those in Table 2, differing from it by 0.01 at most.

What does this imply for the three examples of heterogeneous variances in Section 2? The trials in Cheyne et al. (2008), Adachi et al. (2013), and Santos et al. (2020) each had a balanced design, with 7–10 clusters per treatment arm, and an outcome SD ratio around 2 for several outcomes. From Table 2, it follows that the RE of their designs relative to the MMRED may deviate up to 10% in either direction from the values stated in Sections 6 and 7, depending on the unreported cost ratios in these trials (i.e., up to 10% deviation from a RE of 0.70 if $p = 2$ and of 0.55 if $p = 3$).

9 | DISCUSSION

Optimal sample sizes per level (cluster, individual) of a CRT have been published by Raudenbush (1997) and Moerbeek et al. (2000), based on the assumption of homogeneity between treatments with respect to costs and variances and the additional assumption that the variances are known in the design stage. Unfortunately, the optimal design of a two-arm CRT strongly depends on the four variances involved (one per design level per treatment arm), and misspecification of these variances can lead to an inefficient design. As a solution for this, Van Breukelen and Candel (2018) presented a Maximin design which maximizes the minimum efficiency (MMED) over a range of plausible values for the unknown variances. This design has the advantage of guaranteeing a prespecified power and precision, but the drawback of assuming the largest plausible ICC and thereby requiring a higher study budget and a larger sample size than may be needed.

This paper presented another Maximin design, which maximizes not the minimum efficiency, but the minimum relative efficiency (MMRED), relative to the LODs. The MMRED was compared with the popular balanced design, which is optimal under homogeneity of costs and variances, with the cc design, which is optimal under heterogeneity of costs and homogeneity of variances, and with the MMED. If the costs are homogeneous, then all four designs are balanced irrespective of the heterogeneity of variances. If the costs are heterogeneous while the variances are homogeneous, then both Maximin designs are equal to the cc design. However, under simultaneous heterogeneity of costs and variances, the two Maximin designs behave very differently. Whereas increasing uncertainty about variance heterogeneity moves the MMED away from the cc design towards a balanced design, it moves the MMRED away from the cc design towards a balanced *budget allocation* between both treatment arms, leading to an even more *unbalanced* design than the cc design if the costs are heterogeneous. Further, in terms of the minimum RE criterion, the cc design is always superior to the balanced design unless the costs are homogeneous, in which case the two coincide. This again is different from results based on the minimum efficiency criterion, according to which the cc design beats the balanced design only if the costs are much more heterogeneous than variances (see Van Breukelen & Candel, 2018, for details).

Since the variances are unknown and need to be estimated in the data analysis after trial completion, design comparisons in terms of their relative efficiencies must take into account not only the sampling variance of the treatment effect in each design but also the degrees of freedom of the t -distribution for testing and interval estimation of the treatment effect, as in Equation (16). For a budget large enough to have a total of 40 clusters (20 per arm) in the balanced design, the effect of df on the relative efficiencies of the designs is small for variance and cost ratios up to 9. For a total of 20 clusters (10 per arm), the df effect is small only for variance and cost ratios up to 4.

The present results show that depending on whether we choose to maximize the minimum efficiency (minimize the maximum sampling variance) or to maximize the minimum RE, we may end up with quite different designs: more like the balanced design under the first optimality criterion, more like the cc design under the second criterion. This raises the question of which optimality criterion to choose. The Maximin efficiency criterion is the safest choice in that it guarantees the smallest sampling variance in the worst case of a maximum ICC and a maximum heterogeneity of variance, but it has the drawback of possible overemphasis on an unlikely scenario which requires a large study budget. This can be alleviated by lowering the maximum ICC value and/or by narrowing the range for the SD ratio to exclude extreme scenarios. The Maximin RE criterion is safe in the sense of staying as close to the optimal design as possible over the whole range for the ICC and the SD ratio, thereby avoiding the overemphasis on an unlikely extreme scenario.

However, sample size calculation is easier for the MMED than for the Maximin RE design. For both designs, the user needs to specify the effect size β_1 , significance level α and power $1 - \gamma$, the parameters p and u for the treatment-to-control cost and variance ratios, and the ranges $[V_{\min}, V_{\max}]$ and $[\rho_{\min}, \rho_{\max}]$. For sample size calculation, one furthermore needs to specify an ICC value because $\text{Var}(\hat{\beta}_1)$ is a monotonic function of the ICC (see Equations 3,8,11). For the MMED, the


obvious choice is ρ_{\max} , which is the worst-case scenario, for which the MMED is the LOD. For the MMRED there is no obvious choice, but one might consider choosing that ICC for which the MMRED is the LOD, which is an ICC value in the lower half of the ICC range and reflects a somewhat optimistic scenario. Given this difference between the two Maximin designs, one might consider the safe but pessimistic MMED for the first stage of a group sequential trial which allows early stopping, and the more optimistic MMRED for the first stage of an adaptive design which allows sample size re-estimation, after the first stage. Any sample size calculation should furthermore prevent a too small number of clusters for reliable estimation of the variance components. In practice, this may not often be an issue, however. For example, sample size calculations in Van Breukelen and Candel (2018, table 4) for the MMED, using similar ranges for the ICC, the variance ratio, and the cost ratio as in the present paper always resulted in at least 10 clusters per arm for a medium effect size (Cohen's $d = 0.50$), two-tailed test with size 5% and power 90%. Further, as the authors explained with some references, for this test size and power two clusters need to be added to that in each arm to compensate for the power loss arising from finite df, as their sample size calculation assumed a z -test instead of a t -test.

Finally, the present study is limited to two-arm CRTs with a quantitative outcome and known costs. One logical extension would be to Maximin design (MMED and MMRED) of CRTs with a binary or ordinal outcome or count data or survival times. Another would be to Maximin design of multicenter trials with the center as random effect and center by treatment interaction. A third extension would be to Maximin design of trials with clustering in only one arm, for which the optimal design has been derived by Moerbeek and Wong (2008) under the assumption of a known ICC and a known variance ratio. A fourth and last extension might be a Maximin approach to the costs if these are unknown in the design stage, yet assumed to be heterogeneous.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Gerard J. P. van Breukelen  <https://orcid.org/0000-0003-0949-0272>

Math J. J. M. Candel  <https://orcid.org/0000-0002-2229-1131>

REFERENCES

- Adachi, M., Yamaoka, K., Watanabe, M., Nishikawa, M., Kobayashi, I., Hida, E., & Tango, T. (2013). Effects of lifestyle education program for type 2 diabetes patients in clinics: A cluster randomized controlled trial. *BMC Public Health*, 13, 467. <http://www.biomedcentral.com/1471-2458/13/467>
- Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intracluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57, 785–794. <https://doi.org/10.1016/j.jclinepi.2003.12.013>.
- Berger, M. P. F., & Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*. Wiley
- Candel, M., & Van Breukelen, G. J. P. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research*, 24, 557–573. <https://doi.org/10.1177/0962280214563100>
- Cheyne, H., Hundley, V., Dowding, D., Bland, M., McNamee, P., Greer, I., Styles, M., Barnett, C. A., Scotland, G., & Niven, C. (2008). Effects of algorithm for diagnosis of active labour: Cluster randomized trial. *British Medical Journal*, 337, <https://doi.org/10.1136/bmj.a2396>
- Detle, H., Martinez-Lopez, I., Ortiz-Rodriguez, I. M., & Pepelyshev, A. (2006). Maximin efficient design of experiment for exponential regression models. *Journal of Statistical Planning and Inference*, 136, 4397–4418. <https://doi.org/10.1016/j.jspi.2005.06.006>
- Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. Wiley.
- Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials*, 1, 80–90. <https://doi.org/10.1191/1740774504cn006rr>
- Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68, 155–165. <https://doi.org/10.1037/0022-006X.68.1.155>
- Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomized trials*. Chapman and Hall.

- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87. <https://doi.org/10.3102/0162373707299706>
- Julious, S. A. (2010). *Sample sizes for clinical trials*. Chapman & Hall/CRC.
- King, J., & Wong, W. K. (2000). Minimax D-optimal designs for the logistic model. *Biometrics*, 56, 1263–1267. <http://doi.org/10.1111/j.0006-341X.2000.01263.x>
- Lake, S., Kammann, E., Klar, N., & Betensky, R. (2002). Sample size re-estimation in cluster randomized trials. *Statistics in Medicine*, 21, 1337–1350. <https://doi.org/10.1002/sim.1121>
- Lemme, F., Van Breukelen, G. J. P., & Berger, M. P. F. (2016). Efficient treatment allocation in 2x2 cluster randomized trials when costs and variances are heterogeneous. *Statistics in Medicine*, 35, 4320–4334. <https://doi.org/10.1002/sim.7003>
- Liu, X. (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *Journal of Educational and Behavioral Statistics*, 28, 231–248. <https://doi.org/10.3102/10769986028003231>
- Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational & Behavioral Statistics*, 25, 271–284. <https://doi.org/10.3102/10769986025003271>
- Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicentre intervention studies. *Journal of Clinical Epidemiology*, 56, 341–350. [https://doi.org/10.1016/S0895-4356\(03\)00007-6](https://doi.org/10.1016/S0895-4356(03)00007-6)
- Moerbeek, M., & Wong, W. K. (2008). Sample size formula for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine*, 27, 2850–2864. <https://doi.org/10.1002/sim.3115>
- Müller, C. H. (1995). Maximin efficient designs for estimating nonlinear aspects in linear models. *Journal of Statistical Planning and Inference*, 44, 117–132. [https://doi.org/10.1016/0378-3758\(94\)00042-T](https://doi.org/10.1016/0378-3758(94)00042-T)
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.
- Pronzato, L., & Walter, E. (1988). Robust experiment design via maximin optimization. *Mathematical Biosciences*, 89, 161–176. [https://doi.org/10.1016/0025-5564\(88\)90097-1](https://doi.org/10.1016/0025-5564(88)90097-1)
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173–185. <https://doi.org/10.1037/1082-989X.2.2.173>
- Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2, 152–162. <https://doi.org/10.1191/1740774505cn0760a>
- Rotondi, M. A., & Donner, A. (2009). Sample size estimation in cluster randomized educational trials: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, 34, 229–237. <https://doi.org/10.3102/1076998609332756>
- Santos, H. G., Chiavegato, L. D., Valentim, D. P., & Padula, R. S. (2020). Effectiveness of progressive resistance exercise program for industrial workers during breaks on perceived fatigue control: A cluster randomized controlled trial. *BMC Public Health [Electronic Resource]*, 20, 849. <https://doi.org/10.1186/s12889-020-08994-x>
- Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309–316. <https://doi.org/10.1007/BF02288586>
- Sitter, R. R. (1992). Robust designs for binary data. *Biometrics*, 48, 1145–1155. <https://doi.org/10.2307/2532705>
- Van Breukelen, G. J. P., Candel, M., & Berger, M. P. F. (2008). Relative efficiency of unequal cluster sizes for variance component estimation in cluster randomized and multicentre trials. *Statistical Methods in Medical Research*, 17, 439–458. <https://doi.org/10.1177/0962280206079018>
- Van Breukelen, G. J. P., & Candel, M. (2012). Efficiency loss due to varying cluster size in cluster randomized trials is smaller than literature suggests. *Statistics in Medicine*, 31, 397–400. <https://doi.org/10.1002/sim.4449>
- Van Breukelen, G. J. P., & Candel, M. (2015). Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Statistical Methods in Medical Research*, 24, 540–556. <https://doi.org/10.1177/0962280211421344>
- Van Breukelen, G. J. P., & Candel, M. (2018). Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Statistics in Medicine*, 37, 3027–3046. <https://doi.org/10.1002/sim.7824>
- Van Schie, S., & Moerbeek, M. (2014). Re-estimating sample size in cluster randomized trials with active recruitment within clusters. *Statistics in Medicine*, 33, 3253–3268. <https://doi.org/10.1002/sim.6172>
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362. <https://doi.org/10.2307/2332010>
- Wiens, D. P. (2019). Maximin power designs in testing lack of fit. *Journal of Statistical Planning and Inference*, 199, 311–317. <https://doi.org/10.1016/j.jspi.2018.07.007>
- Wu, S., Wong, W. K., & Crespi, C. M. (2017). Maximin optimal designs for cluster randomized trials. *Biometrics*, 73, 916–926. <https://doi.org/10.1111/biom.12659>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: van Breukelen, G. J. P., & Candel, M. J. J. M. (2021). Maximin design of cluster randomized trials with heterogeneous costs and variances. *Biometrical Journal*, 63, 1444–1463.

<https://doi.org/10.1002/bimj.202100019>

APPENDIX A: LIST OF SYMBOLS USED

| Symbol | Interpretation | Introduced in Section number |
|----------------------------|---|------------------------------|
| β_1 | The treatment effect of interest | 3 |
| σ_u^2 | Residual variance at the cluster level | 3 |
| σ_ε^2 | Residual variance at the individual level | 3 |
| σ_Y^2 | Total residual variance | 3 |
| ρ | Intraclass correlation (ICC) | 3, Equation (2) |
| K | Total number of clusters sampled | 3 |
| n | Number of individuals sampled per cluster | 3 |
| c | Cost per cluster | 3 |
| s | Cost per subject | 3 |
| B | Budget for the study | 3 |
| f | Fraction of the study budget spent on the treated arm | 3 |
| $f/(1-f)$ | Budget allocation ratio | |
| $g_t(\rho_t)$ | $(\sqrt{\rho_t c_t} + \sqrt{(1-\rho_t)s_t})^2$ | 3, Equation (6) |
| u and $1/u$ | Maximum and minimum for the SD ratio $\sigma_{y(t)}/\sigma_{y(c)}$ | 4 |
| V | Shortcut for $\sigma_{y(t)}^2 + \sigma_{y(c)}^2$ | 4 |
| V_{\min}, V_{\max} | Minimum and maximum of V | 4 |
| ρ_{\min}, ρ_{\max} | Minimum and maximum of the ICC | 4 |
| p | $\sqrt{\frac{g_t(\rho_{\max})}{g_c(\rho_{\max})}}$ | 4 |
| $h_t(\rho_t)$ | $[(n_t^r - 1)\rho_t + 1] \left(\frac{c_t + n_t^r s_t}{n_t^r} \right)$ | 5, Equation (12) |
| p_1 | $\frac{\sqrt{h_t(\rho_{\min})}}{\sqrt{h_c(\rho_{\max})}}$ | 5 |
| p_2 | $\frac{\sqrt{h_c(\rho_{\min})}}{\sqrt{h_t(\rho_{\max})}}$ | 5 |
| z | $\frac{\sigma_{y_t} \sqrt{h_t(\rho_t)}}{\sigma_{y_c} \sqrt{h_c(\rho_c)}}$ | Appendix B, Equation (B3) |
| Subscripts | U = random cluster effect; e = random individual effect or residual; y = outcome variable; t = in the treated arm; c = in the control arm | 3 |
| Superscripts | * = optimal design; m = Maximin efficiency (MME) design; r = Maximin relative efficiency (MMRE) design | 3-4-5 |

APPENDIX B: DERIVATION OF THE MMRED

Minimizing Equation (11) as a function of f gives as optimal budget split:

$$\frac{f^*}{1-f^*} = \frac{\sigma_{y(t)}\sqrt{h_t(\rho_t)}}{\sigma_{y(c)}\sqrt{h_c(\rho_c)}} \quad (\text{B1})$$

and gives as optimal (minimum) $\text{Var}(\hat{\beta}_1)$:

$$\text{Var}(\hat{\beta}_1) = \left(\sigma_{y(t)}\sqrt{h_t(\rho_t)} + \sigma_{y(c)}\sqrt{h_c(\rho_c)} \right)^2 / B \quad (\text{B2})$$

which are Equations (7) and (8) apart from replacing $g_t(\rho_t)$ with $h_t(\rho_t)$, and $g_c(\rho_c)$ with $h_c(\rho_c)$, as the sample size per cluster now obeys Equation (10) instead of (4). The optimal budget split in (B1) again depends on the unknown parameter vector $(\sigma_{y(t)}^2, \rho_t, \sigma_{y(c)}^2, \rho_c)$. MMED resolved this dependence by first choosing the parameter values which maximize $\text{Var}(\hat{\beta}_1)$ as given by Equation (5), and then choosing the budget split which minimizes that maximum of $\text{Var}(\hat{\beta}_1)$. MMRED resolves the dependency similarly in the following steps.

First, consider the RE of a given budget ratio $f/(1-f)$ compared to the optimal budget ratio in Equation (B1), which is the ratio of Equation (B2) to Equation (11) and can be written into

$$\text{RE} = \frac{(z+1)^2 f(1-f)}{z^2(1-f)+f}, \text{ where } z = \frac{\sigma_{y(t)}\sqrt{h_t(\rho_t)}}{\sigma_{y(c)}\sqrt{h_c(\rho_c)}}. \quad (\text{B3})$$

Note that z is the optimal budget ratio in Equation (B1). Now, derive the minimum of (B3) as a function of z (unknown parameters), and then find the maximum of that minimum as a function of f (design). To this end, first define

$$p_1 = \frac{\sqrt{h_t(\rho_{\min})}}{\sqrt{h_c(\rho_{\max})}}, p_2 = \frac{\sqrt{h_t(\rho_{\max})}}{\sqrt{h_c(\rho_{\min})}}, \quad (\text{B4})$$

which are the minimum, respectively, maximum, of the factor $\sqrt{h_t(\rho_t)/h_c(\rho_c)}$ in z , since $h_t(\rho_t)$ is an increasing function of ρ_t , and likewise $h_c(\rho_c)$ of ρ_c . Second, note that $\sigma_{y(t)}/\sigma_{y(c)} \in [1/u, u]$ implies that $z \in [p_1/u, p_2u]$.

Taking the partial derivative of (B3) with respect to z , setting it equal to zero, and using the fact that $z, f, (1-f)$ all > 0 gives: $\partial \text{RE} / \partial z = 0$, or $>$, or $<$, $0 \Leftrightarrow z = 0$, or $<$, or $>$, $f/(1-f)$, with maximum RE = 1 at $z = f/(1-f)$, and minimum RE either at $z \rightarrow 0$, giving RE = $(1-f)$, or at $z \rightarrow \infty$, giving RE = f . So, the minimum RE is $\text{Min}(f, 1-f)$, which is maximized by letting $f = 1-f = 0.50$. So, if $z \in (0, \infty)$, then the optimal budget split is 50:50 (giving an unbalanced design due to the heterogeneity of costs).

However, $z \in [p_1/u, p_2u]$, with p_1 and p_2 as defined in Equation (B4). With the same procedure as for unbounded z , we get that the RE is minimized at either of the two boundaries for z , so we only need to consider the RE at either boundary. Taking first the RE at the upper boundary $z = p_2u$, and taking its partial derivative with respect to f shows that the RE at $z = p_2u$ increases from 0 if $f = 0$, to a maximum of 1 if $f/(1-f) = p_2u$, and then decreases back to 0 for $f = 1$. Taking next the RE at the lower boundary $z = p_1/u$, we find that the RE increases from 0 for $f = 0$ to a maximum of 1 for $f/(1-f) = p_1/u$, and then decreases back to 0 for $f = 1$. Since these two REs are single-peaked functions of $f \in [0, 1]$ which attain a maximum at different f -values, or equivalently, at different budget ratios $f/(1-f)$, the minimum of both RE's is maximized by that budget ratio where the two functions intersect, that is, where the RE at $z = p_2u$ and the RE at $z = p_1/u$ are equal. Equating the two REs and rewriting gives as MMRED the following budget split:

$$\frac{fr}{1-fr} = \frac{\left(\frac{p_1}{u}\right)^2 (p_2u+1)^2 - (p_2u)^2 \left(\frac{p_1}{u}+1\right)^2}{\left(\frac{p_1}{u}+1\right)^2 - (p_2u+1)^2}, \quad (\text{B5})$$

By first elaborating numerator and denominator, and then dividing both by $(p_1/u - p_2u)$, which requires the latter to be unequal to zero (and thus rules out the case where $u = 1$ and $\rho_{\min} = \rho_{\max}$ both hold, that is, the case of homogeneous

variance and known ICC), Equation (B5) can be shown to reduce to

$$\frac{f^r}{1 - f^r} = \frac{2p_1p_2 + \left(\frac{p_1}{u}\right) + p_2u}{2 + \left(\frac{p_1}{u}\right) + p_2u}. \tag{B6}$$

Inserting this result in Equation (B3) and using either $z = p_2u$ or $z = p_1/u$ then gives the minimum RE of the MMRED. The budget ratio (B6) goes to 1, and the minimum RE goes to 0.50, as $u \rightarrow \infty$, irrespective of the costs.

The above results are based on a two-sided interval $[1/u, u]$ for the SD ratio. If a researcher knows that the variance will be at least as large in the treated arm as in the control arm, this interval can be replaced with the one-sided interval $[1, u]$. The minimum RE is then obtained by letting $z = p_1$ or $z = p_2u$ in Equation (B3), depending on the budget split. The MMRED budget split can then be shown to be as in Equation (B6), except that p_1/u must be replaced with p_1 in numerator and denominator. The minimum RE of the MMRED is then obtained from Equation (B3) with $z = p_1$ or $z = p_2u$, the two giving the same RE in case of the MMRED. Likewise, if the variance is known to be at most as large in the treated arm as in the control arm, then replace the interval with $[1/u, 1]$ and replace in Equation (B6) p_2u with p_2 in numerator and denominator, and let in Equation (B3) $z = p_2$ or $z = p_1/u$.

APPENDIX C: MINIMUM RELATIVE EFFICIENCIES OF BALANCED AND COST-CONSCIOUS DESIGNS

For comparison with the MMRED in Section 5, this appendix is limited to the case where $c_t/s_t = c_c/s_c$ (homogeneous cost ratio) and $\rho_{\min} = \rho_{\max}$ (known homogeneous ICC), so that $p_1 = p_2 = \sqrt{c_t/c_c} = \sqrt{s_t/s_c}$ (see Equations 12 and 13, which is here denoted by p , for both designs in this appendix, the balanced and the cc design).

The balanced design ($n_t = n_c, K_t = K_c$) needs as budget split $f^b/(1 - f^b) = p^2$. Substituting this in Equation (B3) gives as RE at the two boundaries for z (remember that the RE of any design relative to the LOD is minimized at a boundary for z ; see Appendix B):

$$\text{At } z = pu : \text{RE} = \frac{(pu + 1)^2}{(1 + p^2)(u^2 + 1)} \tag{C1}$$

$$\text{At } z = \frac{p}{u} : \text{RE} = \frac{(p + u)^2}{(1 + p^2)(u^2 + 1)}. \tag{C2}$$

Equation (C1) is the RE of the balanced design versus MMED in terms of the minimum efficiency criterion, for the case $p > u$, and increases in u , but decreases in p , then. Equation (C2) is the RE of the balanced design versus MMED in terms of minimum efficiency for the case $p < u^{-1}$ and increases in u and in p then (Van Breukelen & Candel, 2018, section 5.1 and table 1). So, in both cases the RE of balanced versus MMED in terms of minimum efficiency increases as heterogeneity of variance increases (i.e., as $u \rightarrow \infty$) and as heterogeneity of costs decreases (i.e., as $p \rightarrow 1$). For $p \in [u^{-1}, u]$, the balanced design is the MMED. Of importance here is that the minimum of (C1) and (C2) is the minimum RE of the balanced design versus LOD since that minimum is attained at $z = pu$ or at $z = p/u$ (Appendix B). Taking the minimum of (C1) and (C2) gives (C1) if $p < 1$ (treatment cheaper than control) and (C2) if $p > 1$ (treatment more expensive than control). That minimum RE of the balanced design versus the LOD is plotted in Figure C1 upper panel, and is a decreasing function of u , and $\text{minRE} \rightarrow \min(1, p^2)/(1 + p^2)$ as $u \rightarrow \infty$. Further, the minimum RE of the balanced design increases as $p \rightarrow 1$ for finite u .

The cc design gives as budget split $f^c/(1 - f^c) = p$ by letting $\sigma_{y(t)}/\sigma_{y(c)} = 1$ in Equation (7). Substituting this in Equation (B3) gives after rewriting as RE at the boundaries for z :

$$\text{At } z = pu : \text{RE} = \frac{(pu + 1)^2}{(1 + p)(pu^2 + 1)} \tag{C3}$$

$$\text{At } z = \frac{p}{u} : \text{RE} = \frac{(p + u)^2}{(1 + p)(u^2 + p)}. \tag{C4}$$

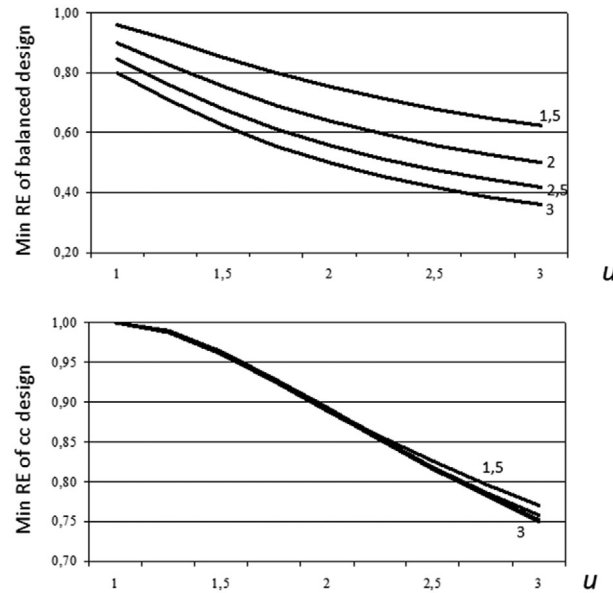


FIGURE C1 Minimum relative efficiency (min RE) of the balanced design (upper panel) and the cc design (lower panel) as a function of the SD ratio range $[u^{-1}, u]$ and p

Equation (C3) is the RE of the cc design versus MMED in terms of the minimum efficiency criterion, for the case $p > u$, and Equation (C4) is the RE of the cc design versus MMED in terms of the same criterion, for the case $p < u^{-1}$ (Van Breukelen & Candel, 2018, section 5.2 and tables 1 and 2). In both cases, the RE of the cc design versus MMED in terms of minimum efficiency *decreases* as heterogeneity of variance *increases* (i.e., as $u \rightarrow \infty$) and as heterogeneity of costs *decreases* (i.e., as $p \rightarrow 1$). These effects are the opposite of those for the balanced versus MMED. Finally, for $p \in [u^{-1}, u]$ the RE of the cc design versus MMED in terms of minimum efficiency is obtained by replacing the numerator in (C3) with the denominator of (C1) if $p \in [1, u]$, and by replacing the numerator in (C4) with the denominator of (C2) if $p \in [u^{-1}, 1]$. In these cases too, the RE of the cc design versus MMED in terms of minimum efficiency *decreases* as $u \rightarrow \infty$, but its behavior as a function of p is more complicated, and if $p = 1$ the cc design is the balanced design and the MMED.

Taking the minimum of (C3) and (C4) gives as minimum RE of the cc design compared to the LOD, respectively, (C3) if $p < 1$ and (C4) if $p > 1$. That minimum RE is plotted in Figure C1 lower panel and is a *decreasing* function of u , and $\text{minRE} \rightarrow \min(1, p)/(1 + p)$ as $u \rightarrow \infty$. Further, the minimum RE of the cc design relative to the LOD increases as $p \rightarrow 1$ for finite u , which is the opposite of the effect of p on the RE of cc versus MMED with respect to the minimum efficiency criterion.

APPENDIX D: DERIVATION OF EQUATION (16)

To obtain Equation (16), first derive the equation for $\text{RE}(D1 \text{ versus } D2)$ if all variance components and thus also the sampling variance of the treatment effect estimator, $\text{Var}(\hat{\beta}_1)$, are known, and define $\text{SE} = \sqrt{\text{Var}(\hat{\beta}_1)}$. The sampling distribution of the test statistic $Z_0 = \hat{\beta}_1/\text{SE}$ is then standard normal under $H_0 : \beta_1 = 0$, and the sampling distribution of $Z_1 = (\hat{\beta}_1 - \Delta)/\text{SE}$ is standard normal under $H_1 : \beta_1 = \Delta$. From this it follows that for any given nonzero value of Δ , the test power is

$$\Phi(z_{1-\gamma}) = P(Z_0 > z_{1-\alpha/2} | H_1) = P\left(Z_1 > z_{1-\alpha/2} - \frac{\Delta}{\text{SE}} | H_1\right) = \Phi\left(-z_{1-\alpha/2} + \frac{\Delta}{\text{SE}}\right), \quad (\text{D1})$$

where $\Phi(\cdot)$ is the standard normal distribution function. Since $\Phi(\cdot)$ is a monotonically increasing continuous function, (D1) implies:

$$z_{1-\gamma} = -z_{1-\alpha/2} + \frac{\Delta}{\text{SE}} \Rightarrow (z_{1-\gamma} + z_{1-\alpha/2})^2 = \frac{\Delta^2}{\text{Var}(\hat{\beta}_1)}. \quad (\text{D2})$$

So, to have the same test power $1 - \gamma$ for the same treatment effect Δ and same test size α designs $D1$ and $D2$ need to have the same $\text{Var}(\hat{\beta}_1)$. Now, $\text{Var}(\hat{\beta}_1)$ is inversely proportional to the study budget B for all four designs in this paper, as in Equations (8) and (B2). To see that this holds for all four designs, note that the sample size per cluster is independent of B in all designs (Equations 4, 10), and the budget ratio likewise (Equations 7,9,14, B1). Therefore, the numbers of clusters per arm are both proportional to B , as in Equation (4). From this and Equation (3) then follows that $\text{Var}(\hat{\beta}_1)$ is inversely proportional to the budget B . Therefore, if $\text{Var}(\hat{\beta}_1|D1) > \text{Var}(\hat{\beta}_1|D2)$, the budget for $D1$ needs to be multiplied with $\text{Var}(\hat{\beta}_1|D1)/\text{Var}(\hat{\beta}_1|D2)$ to make $D1$ as powerful as $D2$, in short: $RE(D1 \text{ versus } D2) = \text{Var}(\hat{\beta}_1|D2)/\text{Var}(\hat{\beta}_1|D1)$.

Now consider the case where all variance components and thus $\text{Var}(\hat{\beta}_1)$ are unknown and estimated so that the test statistics have a Student t -distribution. Equations (D1) and (D2) then apply after replacing $z_{1-\gamma}$ with $t_{df,1-\gamma}$ and $z_{1-\alpha/2}$ with $t_{df,1-\alpha/2}$. It then follows from (D2) that designs $D1$ and $D2$ have the same power for the same Δ and α iff:

$$\frac{\text{Var}(\hat{\beta}_1|D2)}{\text{Var}(\hat{\beta}_1|D1)} \times \frac{(t_{df2,1-\gamma} + t_{df2,1-\alpha/2})^2}{(t_{df1,1-\gamma} + t_{df1,1-\alpha/2})^2} = 1. \quad (\text{D3})$$

Since this implies that the budget of $D1$ needs to be multiplied by the inverse of the left side of Equation (D3) to make $D1$ as powerful as $D2$, the definition of $RE(D1 \text{ versus } D2)$ is as given in Equation (16).