D-optimality of unequal versus equal cluster sizes for mixed effects linear regression analysis of randomized trials with clusters in one treatment arm

Computational Statistics and Data Analysis 54 (2010) 1906–1920

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

D-optimality of unequal versus equal cluster sizes for mixed effects linear regression analysis of randomized trials with clusters in one treatment arm

Math J.J.M. Candel *, Gerard J.P. Van Breukelen

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

ARTICLE INFO

Article history:
Received 10 July 2009
Received in revised form 21 February 2010
Accepted 22 February 2010
Available online 2 March 2010

Keywords:
Asymptotic relative efficiency
Clustering effects of treatments
D-criterion
D_s-criterion
Optimal design
Varying cluster sizes

ABSTRACT

The efficiency loss due to varying cluster sizes in trials where treatments induce clustering of observations in one of the two treatment arms is examined. Such designs may arise when comparing group therapy to a condition with only medication or a condition not involving any kind of treatment. For maximum likelihood estimation in a mixed effects linear regression, asymptotic relative efficiencies (*RE*) of unequal versus equal cluster sizes in terms of the *D*-criterion and D_s-criteria are derived. A Monte Carlo simulation for small sample sizes shows these asymptotic *RE*s to be very accurate for the D_s-criterion of the fixed effects and rather accurate for the *D*-criterion. Taylor approximations of the asymptotic *RE*s turn out to be accurate and can be used to predict the efficiency loss when planning a trial. The *RE* usually will be more than 0.94 and, when planning sample sizes, multiplying both the number of clusters in one arm and the number of persons in the other arm by $1/RE$ is the most cost-efficient way of regaining the efficiency loss.

1. Introduction

Trials evaluating the effect of an intervention are often characterized by observations being correlated within clusters. A well-known case is group or cluster randomized trials (Donner and Klar, 1994; Raudenbush, 1997), where groups (e.g. schools or general practices) are assigned to one of the several treatment conditions. In these designs groups are the units of assignment. Observations may also be clustered when individuals are the units of assignment. This may occur when the treatment itself induces clustering, such as in individually randomized group treatment trials (Pals et al., 2008), where treatments are given to groups of individuals. In such trials interactions between persons within a group may lead to observations being correlated (Bauer et al., 2008; Roberts and Roberts, 2005). It is quite common that the clustering occurs in only one of the treatment arms, such as when group therapy is compared to a condition involving no kind of intervention (e.g. Bauer et al., 2008; Heller-Boersma et al., 2007; Pisinger et al., 2005) or to a condition involving only medication (e.g. Dannon et al., 2004; Haugli et al., 2001).

Even if the treatment is given on an individual basis, instead of groupwise, the treatment may induce clustering. This may occur if several patients are treated by the same therapist. Since it is likely that patients of the same therapist will be treated in a more similar way than patients treated by different therapists (Pals et al., 2008; Roberts, 1999), observations within each therapist will be clustered. Also in this case clustering may occur in only one of the two treatment arms, such as when treatment is contrasted with a waiting-list condition (e.g. Ladouceur et al., 2000; Thompson et al., 1987; Van Minnen et al., 2003) or with a pharmacological or placebo condition (e.g. Jarrett et al., 1999).

* Corresponding author. Tel.: +31 43 388 2273; fax: +31 43 361 8388.
E-mail addresses: math.candel@stat.unimaas.nl (M.J.J.M. Candel), gerard.vbreukelen@stat.unimaas.nl (G.J.P. Van Breukelen).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.02.020

Starting from a particular cost function Moerbeek and Wong (2008) examined optimal designs and derived sample size formulas in case there is clustering in one of two treatment arms. The present study extends their study by examining clusters that are of unequal size. Unequal cluster sizes may be due to variation in actual cluster size, but also due to nonresponse or dropout of subjects, and therefore is a common situation. The efficiency loss due to variation in cluster sizes when focussing on the estimation of the treatment effect has already been examined (Candel and Van Breukelen, 2009). The present study will examine the efficiency loss when considering the ensemble of all model parameters involved. Also the efficiency loss for the subset of all fixed parameters, among which the treatment effect, and the efficiency loss for the subset of all variance components, will be considered. In randomized trials the fixed parameters are usually of primary interest. The standard errors of the fixed effect estimators, however, are a function of the variance components. Furthermore, variance component estimation in itself may be relevant, such as in quality control studies where the variance in health outcomes between clusters (e.g. general practices, therapists or therapy groups) is examined (e.g. Van Berkestein et al., 1999). This motivates studying the efficiency loss also for the variance components. The efficiency criteria that are examined in this paper, are known as the D-criterion and, in case of a subset of parameters, as the D_s-criterion (Atkinson et al., 2007). For each criterion the issue is how much efficiency is lost due to varying cluster sizes and how to compensate for this loss.

In deriving the efficiency loss we assume that the data within each treatment arm are (approximately) normally distributed and are analyzed with mixed effects linear regression. The relative efficiency of unequal versus equal cluster sizes will be derived for the asymptotic case when the model parameters are estimated through maximum likelihood. Furthermore, Taylor approximations of the asymptotic relative efficiencies, that can be of practical use when planning a trial, will be derived. Since in relevant studies (e.g. Calzone et al., 2005; Haugli et al., 2001; Pals et al., 2008; Roberts and Roberts, 2005; Wampold and Serlin, 2000), the number of clusters as well as the cluster sizes themselves are rather small, the asymptotic relative efficiencies and their Taylor approximations will be checked for small samples by an extensive Monte Carlo simulation study, both for maximum likelihood and restricted maximum likelihood estimation. Finally, we will address how to optimally regain the efficiency loss. If we want to minimize the costs involved with a study, should we additionally sample relatively more clusters for one arm or more persons for the other?

The paper is structured as follows. Section 2 presents the mixed effects linear regression model for trials comparing a treatment arm with clustering to a control arm without clustering. In Section 3 the criteria for evaluating the efficiency loss due to varying cluster sizes will be presented. Section 4 will provide explicit expressions for the asymptotic relative efficiencies when comparing equal to unequal cluster sizes, and will also present Taylor approximations for these asymptotic expressions. Section 5 will discuss the design and results of a Monte Carlo simulation that examines the relative efficiency for various cluster size distributions with realistic sample sizes. The accuracy of both the asymptotic relative efficiencies and the Taylor approximations will be discussed. Section 6 explains how to regain the efficiency loss such that the costs of a design are minimized. Section 7 illustrates for an empirical example how to determine sample sizes in case of the D_s-criterion for fixed effects and how to adjust these to repair the efficiency loss that is expected due to varying cluster sizes. The paper closes with some implications for the planning phase of trials.

2. The mixed effects linear regression model

Suppose in a randomized trial cognitive behavioural group therapy is given to patients with a panic disorder, and its effectiveness in terms of anxiety reduction is compared to only receiving medication (e.g. paroxetine) (cf. Dannon et al., 2004). Within the treatment arm we then have K therapy groups, or, more generally, K clusters. In cluster j ($j = 1, \ldots, K$) there are n_j persons, with all persons receiving the treatment. The total number of persons in the treatment arm therefore amounts to $N = \sum_{j=1}^{K} n_j$. In the control arm medication is given and there is no clustering. This condition is denoted as the $K+1$th cluster, consisting of n_{K+1} persons. If the cluster sizes are equal, we have $n_j = n$ for $j = 1, \ldots, K$, but in general not for $j = K+1$ (e.g. when $n_{K+1} = N$).

Let the outcome variable be some quantitative measure of anxiety, such as the Hamilton rating scale for anxiety (cf. Dannon et al., 2004), which is denoted as y_{ij}, for person i in cluster j ($j = 1, \ldots, K+1$). If y_{ij} is (approximately) normally distributed, mixed effects linear regression is an appropriate tool for data analysis. The corresponding analysis model is then as follows (cf. Bauer et al., 2008; Moerbeek and Wong, 2008; Roberts, 1999):

$$y_{ij} = \beta_0 + (\beta_1 + u_{0j} + \varepsilon_{ij})T_{ij} + \delta_{ij}(1 - T_{ij}), \tag{1}$$

where T_{ij} denotes the treatment condition for person i in cluster j, and is coded as 1 for persons in the treatment arm and 0 for persons in the control arm. With this coding scheme, β_0 represents the mean anxiety score of the control condition (i.e. pharmacological treatment) and β_1 represents the treatment effect of group therapy versus medication on anxiety. The terms ε_{ij} and u_{0j} represent a random person and random cluster effect in the treatment arm, which are assumed to be independently normally distributed with variances σ_ε^2 and σ_0^2 respectively. The random person effect in the control arm, δ_{ij}, is also independently normally distributed, with a possibly different variance σ_δ^2. So the model has five parameters that have to be estimated: two fixed regression weights, β_0 and β_1, and three variance components, σ_0^2, σ_ε^2 and σ_δ^2. Estimates of these parameters can be obtained through maximum likelihood (ML). A relevant concept is the intraclass correlation, which is the correlation between outcome measures for two randomly drawn persons from the same cluster in the treatment arm. The intraclass correlation, denoted as ρ, can be expressed in terms of the variance components as: $\rho = \sigma_0^2/(\sigma_0^2 + \sigma_\varepsilon^2)$.

3. Efficiency criteria

A commonly used criterion for evaluating estimators of parameters is the determinant of the variance–covariance matrix of these estimators (see e.g. Kessels et al., 2008; Moerbeek, 2005; Ortega-Azurduy et al., 2008; Tekle et al., 2008). This is known as the generalized variance of the parameter estimators and denoted as the *D*-criterion (Atkinson et al., 2007). Although the *D*-criterion originally was defined for the fixed parameters of the model (Atkinson et al., 2007), in this study, similar to Candel et al. (2008), we will consider the determinant of all model parameters involved, also including the variance components. The determinant of a subset of the parameters is denoted as the D_s-criterion (cf. Atkinson et al., 2007), and more specifically as $D_s(fixed)$ for the fixed parameters and as $D_s(random)$ for the variance components. As explained in the introduction, variance components are also considered since they are part of the standard errors of the fixed effect estimators, and since their estimation is important in, for instance, quality control studies. Optimal designs that minimize these three criteria, that is, the *D*-criterion, $D_s(fixed)$ or $D_s(random)$, can be shown to be invariant under linear transformations of the independent model variables (see Appendix A for a proof). Moreover, designs that minimize $D_s(fixed)$ have been shown to minimize the area of the confidence region for the fixed parameters (Anderson, 1958). To examine the efficiency loss due to varying cluster sizes, we compare unequal to equal cluster sizes for each of these three criteria.

Let $\boldsymbol{\theta}_f$ denote the vector of the f fixed parameters, let $\boldsymbol{\theta}_r$ denote the vector of r variance components. Then $\boldsymbol{\theta}^T = [\boldsymbol{\theta}_f^T, \boldsymbol{\theta}_r^T]$ is the vector of all, that is $p = f + r$, model parameters. For the model in Eq. (1), we have $p = 2 + 3 = 5$. Let ξ denote the design of a study. The variance–covariance matrix of the estimators $\hat{\boldsymbol{\theta}}$ given a design ξ is denoted as $\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid \xi)$. Let $\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid \xi))$ denote the determinant of this variance–covariance matrix of the estimators $\hat{\boldsymbol{\theta}}$. In what follows, ξ^* denotes a design with equal cluster sizes: $n_j = n$ for $j = 1, \ldots, K$, but not necessarily $n_{K+1} = n$. Further, ξ denotes a design with unequal cluster sizes, but with the same number of clusters, K, the same sample size N of the treatment arm and the same sample size n_{K+1} in the control arm as ξ^*. The relative efficiency (RE) of design ξ compared to design ξ^* in terms of the *D*-criterion is:

$$RE(D) = \left(\frac{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid \xi^*))}{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid \xi))} \right)^{1/p}. \tag{2}$$

The RE in terms of the D_s-criterion can be defined for the fixed parameters, $\boldsymbol{\theta}_f$, and for the variance components, $\boldsymbol{\theta}_r$, as follows respectively:

$$RE(D_s(fixed)) = \left(\frac{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}}_f \mid \xi^*))}{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}}_f \mid \xi))} \right)^{1/f} \quad \text{and} \quad RE(D_s(random)) = \left(\frac{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}}_r \mid \xi^*))}{\mathrm{Det}(\mathrm{Var}(\hat{\boldsymbol{\theta}}_r \mid \xi))} \right)^{1/r}. \tag{3}$$

Since asymptotically there are no correlations between ML estimators of the fixed parameters and the variance components (see e.g. McCulloch and Searle, 2001, p. 176), for ML estimation the relation between the relative efficiencies for the *D*-criterion and the D_s-criteria can be shown to be as follows:

$$RE(D) = [RE(D_s(fixed))]^{2/5} \times [RE(D_s(random))]^{3/5}. \tag{4}$$

For given relative efficiencies in terms of $D_s(fixed)$ and $D_s(random)$, Eq. (4) yields the relative efficiencies in terms of the *D*-criterion. In what follows we will therefore focus on $D_s(fixed)$ and $D_s(random)$.

4. Asymptotic relative efficiencies and their Taylor approximations

To present the asymptotic relative efficiencies of the ML estimators in terms of the D_s-criteria, we need some further notation. Let $\bar{n}$ denote the average cluster size of the K clusters in the treatment arm and let w_j be defined as: $w_j = (\sigma_0^2 + \sigma_\varepsilon^2/n_j)^{-1}$. For equal cluster sizes, we have $n_j = \bar{n}, j = 1, \ldots, K$, and the weight w_j is denoted as w_e. The expressions for the relative efficiencies (RE) based on the D_s-criteria can be shown to be as follows (see Appendices B and C for proofs):

$$RE(D_s(fixed)) = \sqrt{\frac{\sum_{j=1}^{K} w_j}{K w_e}} = \sqrt{\frac{\bar{n} + (1-\rho)/\rho}{\bar{n}} \times \frac{1}{K} \sum_{j=1}^{K} \left(\frac{n_j}{n_j + (1-\rho)/\rho} \right)}, \quad \text{and} \tag{5}$$

$$RE\,(D_s(random)) = \left(\frac{N \sum_{j=1}^{K} w_j^2 - \left(\sum_{j=1}^{K} w_j \right)^2}{(N-K) K w_e^2} \right)^{1/3}. \tag{6}$$

The following properties hold for $RE(D_s(fixed))$ (see Eq. (5)):

1. The relative efficiencies do not depend on the number of clusters K, but do depend on the intraclass correlation ρ and the distribution of cluster sizes.
2. When we multiply each n_j by a factor $c > 0$ and $(1-\rho)/\rho$ by the same factor, this does not affect the relative efficiencies. As a consequence, the minimum of the RE will not depend on $\bar{n}$, but will only be achieved at another value of ρ.
3. When $\rho \to 0$ or $\rho \to 1$, we have $RE \to 1$. For $0 < \rho < 1$, we can show that the $RE \leq 1$ (as a result of the Jensen inequality; see Mood et al., 1974, p. 72), and $RE = 1$ if and only if n_j is constant (and thus $n_j = \bar{n}$), implying that equal cluster sizes are most efficient.

Note that if β_0 in Eq. (1) is known from other studies, it need not be estimated and only one treatment arm is needed in a trial. In this case the one-variance component model (Demidenko, 2004, p. 66) results and the relative efficiency for the single fixed parameter can be shown to be the same as the relative efficiency in terms of $D_s(fixed)$ of a linear mixed model for cluster randomized trials. The latter has already been studied extensively (Candel et al., 2008; Van Breukelen et al., 2007).

For $D_s(random)$ in Eq. (6), the relative efficiencies also satisfy property 1. Property 2 holds for $\bar{n}$ large enough such that $N/(N-K) \approx 1$. Property 3 does not hold, in that, when $\rho \to 0$, the asymptotic RE may become larger than 1. However when $\rho \to 1, RE \to 1$. That asymptotic REs may become larger than 1 for the variance component estimators, was found for cluster randomized trials and multicentre trials (Van Breukelen et al., 2008). A Monte Carlo simulation study also showed this to be the case for the simulated REs for small samples (Candel et al., 2008).

The relative efficiencies relate to those of cluster randomized trials (Van Breukelen et al., 2007, 2008) as follows. When the expression for $RE(D_s(fixed))$ in Eq. (5) is squared, it is equal to $RE(D_s(fixed))$ for cluster randomized trials. When the expression for $RE(D_s(random))$ in Eq. (6) is taken to the power 1.5, it is equal to $RE(D_s(random))$ for cluster randomized trials. This implies that the REs in terms of $D_s(fixed)$ and $D_s(random)$ are closer to one than the corresponding REs for cluster randomized trials. So, asymptotically, the efficiency loss for a trial where clustering only occurs in one of the treatment arms is equal to or smaller than the efficiency loss for trials where clustering occurs in both treatment arms. The intuitive explanation is that with clustering in one of the treatment arms, variation in cluster sizes will only affect the statistical information in one of the arms, and thus the efficiency loss will be smaller compared to when there is clustering in both arms.

For planning the sample sizes of a study it will be useful to have an approximation of the relative efficiency without having to specify the exact distribution of cluster sizes. For cluster randomized trials second-order Taylor approximations have been derived for the REs in Eq. (3) (see Van Breukelen et al., 2007, 2008). Since the REs for cluster randomized trials are simply related to the REs for the partially clustered designs that we consider here, Taylor approximations for Eqs. (5) and (6) immediately follow from these studies.

Consider n_j in Eqs. (5) and (6) for $j = 1, \ldots, K$ as independent realizations of a random variable with expectation μ_n and standard deviation σ_n. Let $CV = \sigma_n/\mu_n$ be the coefficient of variation of cluster sizes. Furthermore, let $\lambda = (\mu_n/(\mu_n + \alpha))$ with $\alpha = \sigma_\varepsilon^2/\sigma_0^2 = (1-\rho)/\rho$. The following second-order Taylor approximation of the RE in Eq. (5) can be given (cf. Van Breukelen et al., 2007):

$$RE_T(D_s(fixed)) = \sqrt{1 - CV^2\lambda(1-\lambda)}. \tag{7}$$

According to this approximation the relative efficiency depends on the coefficient of variation of the cluster sizes, CV, and, through λ, on the mean cluster size and the intraclass correlation. If $\rho \to 0$ or $\rho \to 1$, then $RE_T(D_s(fixed)) \to 1$. For $0 < \rho < 1$ it holds that $RE_T(D_s(fixed)) \leq 1$, with its minimum $\sqrt{1 - \frac{CV^2}{4}}$ being achieved at $\rho = 1/(\mu_n + 1)$. Furthermore, $RE_T(D_s(fixed))$ decreases as CV increases.

For Eq. (6) the second-order Taylor approximation also follows from previous results (Van Breukelen et al., 2008):

$$RE_T(D_s(random)) = \left(1 + CV^2(1-\lambda)(1-3\lambda)\right)^{1/3}. \tag{8}$$

From Eq. (8) it follows that as $\rho \to 0$, that $RE_T(D_s(random)) \to (1 + CV^2)^{1/3}$, which is its maximum, and as $\rho \to 1$ that $RE_T(D_s(random)) \to 1$. The minimum value of $RE_T(D_s(random))$ is $\left(1 - \frac{CV^2}{3}\right)^{1/3}$ and occurs at $\rho = \frac{2}{\mu_n + 2}$. Also, when CV increases the minimum of $RE_T(D_s(random))$ decreases.

Finally, the second-order Taylor expression for the D-criterion follows from Eqs. (4), (7) and (8):

$$RE_T(D) = \left(1 - CV^2\lambda(1-\lambda)\right)^{1/5}\left(1 + CV^2(1-\lambda)(1-3\lambda)\right)^{1/5}. \tag{9}$$

From Eq. (9) it follows that if $\rho \to 0$, then $RE_T(D) \to (1+CV^2)^{1/5}$, which is its maximum, and if $\rho \to 1$ then $RE_T(D) \to 1$. The minimum of $RE_T(D)$ turns out to have a rather complicated expression. However, its minimum is higher than the minimum of the minima of $RE_T(D_s(fixed))$ and $RE_T(D_s(random))$, and, since the minimum of $RE_T(D_s(fixed))$ is the smallest, when interest is in the D-criterion, taking the minimum of $RE_T(D_s(fixed))$ in planning a trial would be a safe strategy. As can be shown by numerical evaluation, taking the minimum of $RE_T(D_s(fixed))$ also is nearly correct, since, for $CV \leq 1$, the difference between the minima of $RE_T(D_s(fixed))$ and $RE_T(D)$ is smaller than 0.01.

Table 1
Overview of the conditions of the Monte Carlo simulation.

Factor	Levels
Distribution of cluster sizes[a]	Unimodal, uniform, bimodal, positively skewed, negatively skewed distribution
Range of cluster sizes[b]	4, 12
Intraclass correlation	$\rho = 0.01$ up to 0.30, with steps of 0.01
Estimation method	ML, REML

[a] More details on the different cluster size distributions are given in Table 2.
[b] The exception is the negatively skewed distribution. Similar to other distributions, for $\bar{n} = 6$ the range is 4, but, since we require cluster sizes ≥ 3, for $\bar{n} = 10$ the range is 8 instead of 12.

5. Monte Carlo investigation of the relative efficiencies

We will examine to what extent the asymptotic results on *RE* as well as the corresponding Taylor approximations hold for numbers of clusters and cluster sizes that are representative of realistic samples through an extensive Monte Carlo simulation study. In the context of comparing group to individual treatments, the number of clusters as well as the cluster sizes are typically small (e.g. Calzone et al., 2005; Haugli et al., 2001; Pals et al., 2008; Roberts, 1999). This also holds for clinical trials where clustering is induced by the therapist (e.g. Jarrett et al., 1999; Ladouceur et al., 2000; Wampold and Serlin, 2000).

5.1. Design of the simulation study

For all simulations we assumed 50%–50% allocation, that is, 50% of the subjects are assigned to the treatment arm and 50% to the control arm. The error variance for the treatment and the control arm (σ_ε^2 and σ_δ^2) were taken to be equal. Other allocation ratios and variance ratios were also examined, but gave similar results and will therefore not be discussed here. The following factors influence the asymptotic relative efficiencies and were systematically varied: (1) the frequency distribution of the cluster sizes, (2) the range of cluster sizes, (3) the size of the intraclass correlation, ρ, and (4) the estimation method. Table 1 displays the choices made for these factors, the motivation is given in what follows.

5.1.1. Frequency distribution

Five different cluster size distributions were studied: (1) a unimodal, (2) a uniform, (3) a bimodal, (4) a positively skewed and (5) a negatively skewed distribution. Three different cluster sizes, g_a, g_b, g_c, with respective frequencies f_a, f_b and f_c were employed. Details of the cluster size distributions can be found in Table 2.

5.1.2. Range of the cluster sizes

We chose an average cluster size $\bar{n} = 6$, with a range of 4. Since larger ranges of cluster sizes are not realistic for such small cluster sizes, also a larger average cluster size of $\bar{n} = 10$, was examined, the range of cluster sizes then being 12 for most distributions. Since we are interested in how well the asymptotic relative efficiencies describe the efficiency loss, and this is especially relevant for large losses, rather extreme ranges were chosen. The average cluster sizes are representative of cluster sizes commonly encountered in trials comparing group to individual interventions (e.g. Bauer et al., 2008; Calzone et al., 2005; Dannon et al., 2004; Haugli et al., 2001; Heller-Boersma et al., 2007; Pals et al., 2008; Roberts, 1999), and in trials with therapist-induced clustering in one of the treatments arms (e.g. Jarrett et al., 1999; Roberts, 1999; Thompson et al., 1987; Wampold and Serlin, 2000).

The number of clusters was fixed at $K = 12$. As was shown in Section 4, the asymptotic relative efficiencies do not depend on K, and choosing a larger $\bar{n}$ gives the same *RE* at a smaller value of the intraclass correlation ρ. However, K and $\bar{n}$ were deliberately chosen to be small to check the accuracy of the asymptotic *RE*s.

5.1.3. Intraclass correlation

The intercept variance σ_0^2 varied from 1 to 30, with the error variance in the treatment condition σ_ε^2 simultaneously varying from 99 to 70, to keep the total variance in the treatment arm at 100. As a result, the intraclass correlation in the treatment arm, ρ, varied by steps of size 0.01 between 0.01 and 0.30, which represents the range of intraclass correlations as commonly encountered in cross-sectional studies (Parker et al., 2005; Smeeth and Siu-Woon, 2002).

5.1.4. Estimation method

Two different estimation methods were considered: Maximum Likelihood (ML) estimation and REstricted Maximum Likelihood (REML) estimation. For both methods negative estimates of the variance components were truncated to 0.

Table 2
Distributions of the cluster sizes in the treatment arm as examined in the Monte Carlo simulation.

Distribution[a,b,c]	Cluster sizes			Range of cluster sizes	*CV*
	g_a	g_b	g_c		
Uniform ($f_a = 4, f_b = 4, f_c = 4$)	4	10	16	12	0.49
	4	6	8	4	0.27
Unimodal ($f_a = 3, f_b = 6, f_c = 3$)	4	10	16	12	0.42
	4	6	8	4	0.24
Bimodal ($f_a = 5, f_b = 2, f_c = 5$)	4	10	16	12	0.55
	4	6	8	4	0.30
Positively skewed ($f_a = 6, f_b = 4, f_c = 2$)	7	10	19	12	0.42
	5	6	9	4	0.24
Negatively skewed ($f_a = 2, f_b = 4, f_c = 6$)	4	10	12	8	0.28
	3	6	7	4	0.24

[a] f_a = number of clusters of size g_a (small), f_b = number of clusters of size g_b (medium), f_c = number of clusters of size g_c (large).
[b] Cluster sizes and cluster frequencies are chosen such that the total number of clusters in the treatment arm is equal to $K = 12$, the average cluster size $\overline{n} = 6$ or $\overline{n} = 10$, g_b is equal to the average cluster size, the smallest group size $g_a \geq 3$ and the range of the cluster sizes ($=g_c - g_a$) varies between 12 and 4.
[c] *CV* = coefficient of variation of the cluster size distribution.

5.2. Simulation procedure

For each of the 300 simulation conditions (=5 distributions × 2 ranges of cluster sizes × 30 intraclass correlations) 10,000 data sets were generated. Each data set represents the data for 12 clusters in the treatment arm consisting, on the average, of 6 or 10 persons and for the control arm consisting of either 72 or 120 persons respectively. The simulations as well as the estimation of the model parameters were performed in version 1.10.0007 of MLwiN (Rasbash et al., 2000). To obtain ML and REML estimates of the parameters, the "Iterative Generalized Least Squares" and "Restricted Iterative Generalized Least Squares" algorithms were employed. In model estimation, the convergence criterion was set to 0.001 and there was no limit on the number of iterations. From the obtained estimates, the *RE*s as defined in Section 3, were calculated.

5.3. Results on the asymptotic relative efficiency and Taylor approximation

Results are shown for $\overline{n} = 10$ and a range of 12. In Fig. 1 the relative efficiencies in terms of the *D*-criterion are displayed. Cluster size distributions are shown that give the highest and the lowest values for the minimum *RE*: a unimodal and a positively skewed distribution on the one hand and a bimodal distribution on the other. For the unimodal and positively skewed distribution, the asymptotic *RE*s are rather close to the simulated *RE*s (discrepancy <2%). For the bimodal distribution the discrepancy between asymptotic and simulated *RE*s is larger and for small intraclass correlations ($\rho < 0.05$) can become as large as 4%. For most distributions the asymptotic *RE*s overestimate the simulated *RE*s. In case of the bimodal distribution the minimum *RE* approaches 0.94. For most distributions however, similar to the results for the unimodal distribution, the relative efficiency exceeds 0.96. For a positively and a negatively skewed distribution (the latter is not shown) the minimum *RE* even exceeds 0.97.

For a smaller cluster size $\overline{n} = 6$, the asymptotic *RE*s generally are somewhat closer to the simulated *RE*s (discrepancy <1.5% for most distributions). Since the coefficient of variation was smaller for $\overline{n} = 6$, this indicates that the asymptotic *RE* is closer to the simulated *RE* in case the *CV* is lower. Also here for most distributions the asymptotic *RE*s overestimate the simulated *RE*s. Since the coefficient of variation was smaller, the *RE*s were larger. Table 3 gives an overview of minimum values for the *RE*, also for the uniform distribution, for which the minimum *RE* takes an intermediate position between the bimodal and unimodal distribution. For $\overline{n} = 6$, the minimum *RE*s for skewed distributions (not shown) are close to those for the unimodal and uniform distribution.

The relative efficiencies in terms of $D_s(fixed)$ are shown in Fig. 2. For the fixed effects, the asymptotic *RE*s describe the simulated *RE*s very adequately (discrepancy <1%). In the extreme case of the bimodal distribution, the *RE* still exceeds 0.94. For the other distributions, similar to the unimodal distribution, the *RE* always exceeds 0.96. In case of a positively and negatively skewed distribution (not shown) the *RE* even exceeds 0.97.

For the relative efficiencies in terms of $D_s(random)$, Fig. 3 shows that the asymptotic *RE*s describe the simulated *RE*s less adequately: the discrepancy may become as large as 3% (and for *RE*s larger than 1 even more than 3%). For the bimodal distribution, the *RE* always exceeds 0.94, for the other distributions it exceeds 0.95.

Also for $\overline{n} = 6$ the asymptotic *RE*s are very close to the simulated *RE*s in case of fixed effects (discrepancy <1%), while the discrepancy may be larger for variance components (up to 3%). An overview of the minimum values for *RE* in case $\overline{n} = 6$, is given in Table 3. For this smaller cluster size, the skewed distributions have minimum *RE*s close to those for the unimodal distribution.

If $\overline{n} = 10$ and the *RE* is defined in terms of the *D*-criterion or $D_s(random)$, the ML estimator has a somewhat higher *RE* than the REML estimator. If $\overline{n} = 10$ and the *RE* is defined in terms of $D_s(fixed)$, or if $\overline{n} = 6$, then there are hardly any differences between both estimation methods.

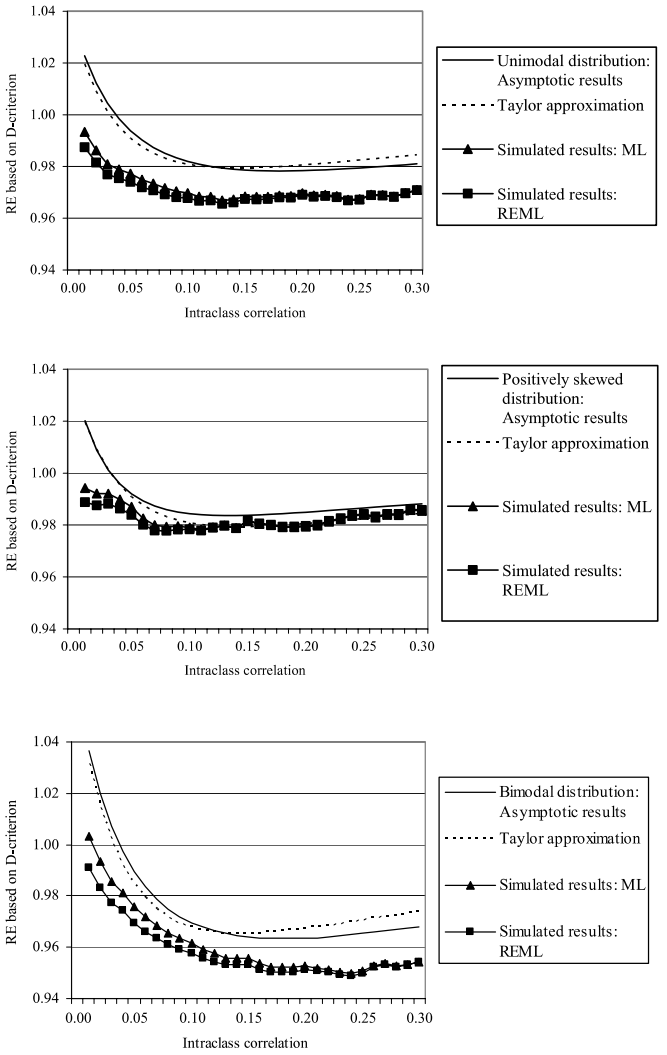

Fig. 1. The relative *D*-criterion for a unimodal (top), a positively skewed (middle) and a bimodal (bottom) distribution of cluster sizes, with $\overline{n} = 10$ and $g_c - g_a = 12$. Displayed is also the asymptotic relative *D*-criterion for the ML estimators and its second-order Taylor approximation.

Table 3
Overview of the minimum relative efficiencies based on the Monte Carlo simulations for ML and REML, specified for the bimodal, uniform and unimodal distribution of cluster sizes, both for a range of the cluster sizes (=$g_c - g_a$) 4 and 12.

Features of the cluster size distribution	$K = 12, \overline{n} = 6, g_c - g_a = 4$			$K = 12, \overline{n} = 10, g_c - g_a = 12$		
Efficiency criterion	Bimodal	Uniform	Unimodal	Bimodal	Uniform	Unimodal
D	0.97	0.98	0.98	0.94	0.96	0.96
$D_s(fixed)$	0.98	0.98	0.98	0.94	0.95	0.96
$D_s(random)$	0.97	0.97	0.98	0.94	0.95	0.96

Comparing the same distributions for small and large ranges of cluster sizes (and thus for small and large coefficients of variation) in Table 3, shows that the simulated *RE*s, in line with the Taylor approximations, have a lower minimum value for larger ranges of cluster sizes. As illustrated by Figs. 1–3, the Taylor approximations are close to the asymptotic *RE*s. This also is true for distributions that are not displayed. In cases that the asymptotic *RE*s are close to the simulated *RE*s, the Taylor approximations therefore are useful in planning a trial.

6. Regaining the efficiency loss due to varying cluster sizes

From Eq. (3) it follows that the efficiency loss for the D_s-criteria can be restored by changing the sample sizes such that $D_s(fixed)$ and $D_s(random)$ for unequal group sizes become $RE(D_s(fixed))^f$ and $RE(D_s(random))^r$ as large respectively.

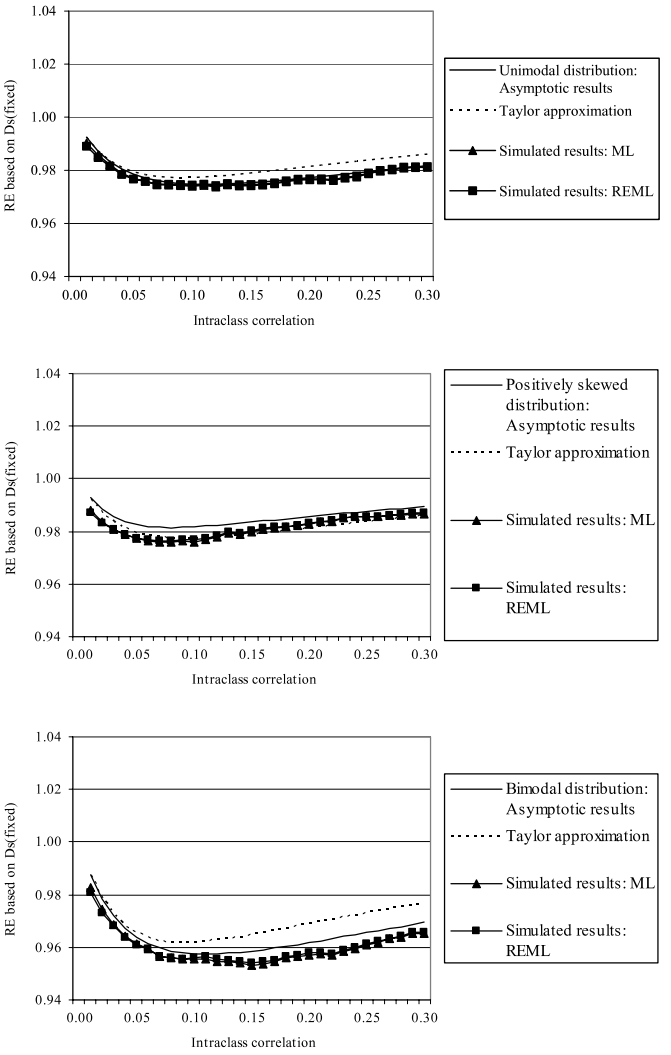

Fig. 2. The relative D_s-criterion of the fixed parameters for a unimodal (top), a positively skewed (middle) and a bimodal (bottom) distribution of cluster sizes, with $\overline{n} = 10$ and $g_c - g_a = 12$. Displayed is also the asymptotic relative D_s-criterion for the ML estimators and its second-order Taylor approximation.

By Eq. (4) this also implies restoring the efficiency loss in terms of the D-criterion. Note that for the model under study (see Eq. (1)), we have $f = 2$ and $r = 3$.

For the three efficiency criteria, let RE be the shorthand notation of the relative efficiencies $RE(D_s(fixed))$, $RE(D_s(random))$ or $RE(D)$, depending on the context. Let f_1 and f_2 denote the replication factors of the number of persons in the control arm and the number of groups in the treatment arm respectively. We will see that in case one wants to minimize the costs involved in a study, $f_1 = f_2 = 1/RE$ is the optimal choice for all three efficiency criteria. So n_{K+1} and K have to be multiplied by the same factor.

We first define a cost function. Let c_t be the costs for each person in the treatment condition, and let c_c be the costs attached to each person in the control condition. Furthermore, let c_g be the extra costs attached to each of the groups in the treatment condition. The total costs of the design, C, can then be given as:

$$C = n_{K+1}c_c + \overline{n}Kc_t + Kc_g. \tag{10}$$

This cost function can also be used if one is interested in minimizing the total number of subjects, that is $n_{K+1} + \overline{n}K$, simply by setting $c_t = c_c = 1$ and $c_g = 0$ in Eq. (10). Since often there is an ideal group size for the treatment condition, we may consider $\overline{n}$ fixed and Eq. (10) can be rewritten as:

$$C = n_{K+1}c_c + K(\overline{n}c_t + c_g) = n_{K+1}c_c + Kc_t^*, \tag{11}$$

where c_t^* denotes $\overline{n}c_t + c_g$, the costs of one group in the treatment condition.

When compensating for the efficiency loss due to varying group sizes in the treatment condition, one would like to replicate the design, minimizing this cost function. Appendix D shows that for each of the D-criteria the efficiency loss can

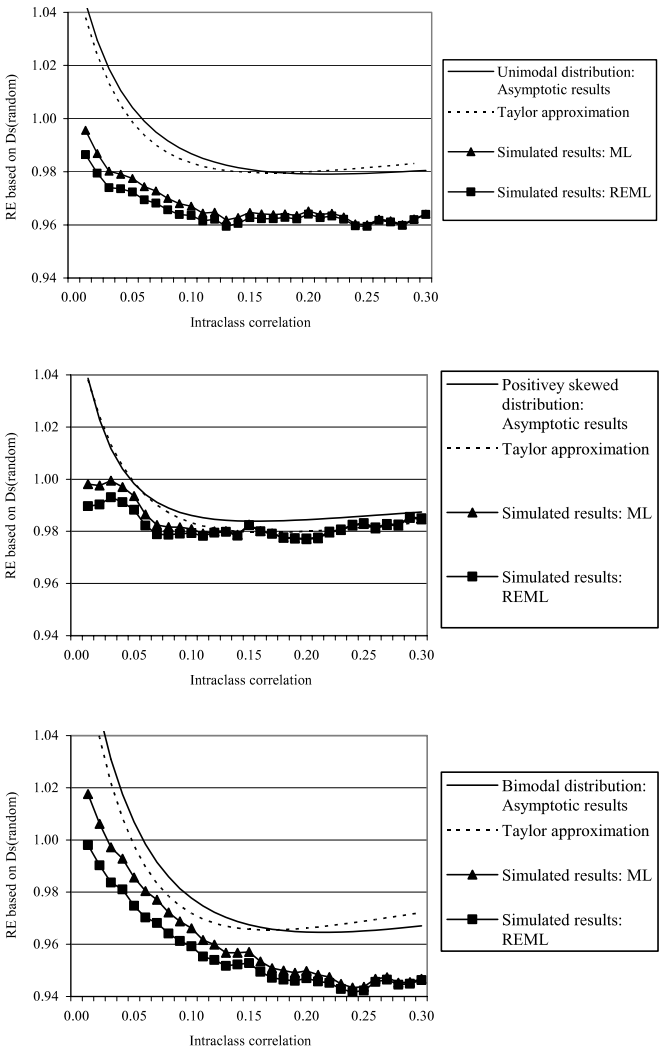

Fig. 3. The relative D_s-criterion of the variance components for a unimodal (top), a positively skewed (middle) and a bimodal (bottom) distribution of cluster sizes, with $\overline{n} = 10$ and $g_c - g_a = 12$. Displayed is also the asymptotic relative D_s-criterion for the ML estimators and its second-order Taylor approximation.

be restored at the lowest costs by replicating the number of groups and the number of persons in the control condition each $1/RE$ times (i.e. $f_1 = f_2 = 1/RE$). Although these replication factors are not optimal for restoring the efficiency loss in terms of $\text{var}(\hat{\beta}_1)$, they have been shown to be highly efficient for that criterion, in that costs differ from the costs for the optimal replication factors by less than 1% in almost all cases (Candel and Van Breukelen, 2009).

7. Empirical illustration

We will first illustrate how to calculate the number of clusters, K, and the size of the ungrouped condition, n_{K+1}, for a trial where the cluster sizes are equal. Next, we will show how to adapt K and n_{K+1} to compensate for the efficiency loss resulting from varying cluster sizes.

As an example, suppose we would like to replicate the Reconnecting Youth prevention programme discussed by Bauer et al. (2008). The programme is meant for high-risk adolescents, defined as those adolescents who had a low grade point average and a high level of truancy. In the treatment arm this programme is given to groups of adolescents, aiming, amongst others, at the reduction of deviant peer bonding. In the control condition (ungrouped) high-risk adolescents will receive no treatment. In estimating the mean post-treatment levels of deviant peer bonding for each of the two arms, a certain level of precision is required. For this purpose, we can reformulate the analysis model in Eq. (1) as:

$$y_{ij} = (\beta_0^* + \delta_{ij})(1 - T_{ij}) + (\beta_1^* + u_{0j} + \varepsilon_{ij})T_{ij}, \tag{12}$$

where T_{ij} denotes the treatment condition for person i in cluster j, and is coded as 1 for persons in the treatment arm and 0 for persons in the control arm. In this alternative formulation, β_0^* represents the mean score of the control condition and β_1^* the

mean score of the treatment condition. Asymptotically, the ML estimators of β_0^* and β_1^* in Eq. (12) are normally distributed and the surface of their $100(1-\alpha)\%$ confidence ellipse, S, is given by (Johnson and Wichern, 2002, p. 127):

$$S = \pi \chi^2_{df=2;\alpha} D_s(fixed)^{1/2}, \tag{13}$$

where $\chi^2_{df=2;\alpha}$ is the $100(1-\alpha)\%$ percentile of a chi-square distribution with two degrees of freedom.

In determining the sample size, we require the surface of the $100(1-\alpha)\%$ confidence ellipse not to exceed some criterion value (cf. Rosner, 2006, p. 259). The reparameterization in Eq. (12) simplifies the derivation of sample sizes, in that, as we will see, it simplifies establishing a criterion value for the surface of the confidence ellipse of the fixed parameters. It can be shown that Eq. (B.2) for $D_s(fixed)$ holds for the model in Eq. (12) as well as for the model in Eq. (1) (see Appendices A and B for a proof). As a consequence, the reparameterization in Eq. (12) does not affect the asymptotic relative efficiency, and, by Eq. (13), also not the surface of the confidence ellipse and hence the calculation of sample sizes.

For the case of equal cluster sizes Eq. (B.2) reduces to

$$D_s(fixed) = \frac{\sigma_\delta^2}{n_{K+1}} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{K}, \tag{14}$$

yielding the following expression for the surface of the $100(1-\alpha)\%$ confidence ellipse:

$$S = \pi \chi^2_{df=2;\alpha} \left(\frac{\sigma_\delta^2}{n_{K+1}} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{K} \right)^{1/2}. \tag{15}$$

Since the ML estimators of β_0^* and β_1^* in Eq. (12) (as opposed to the ML estimators of β_0 and β_1 in Eq. (1)) are independent (compare Eqs. (B.1) and (B.3)), the criterion value for the surface in Eq. (15) is proportional to the maximum width of the confidence ellipse for the population mean in the treatment condition, L_t, times the maximum width of the confidence ellipse for the population mean in the control condition, L_c. More precisely, the criterion value can be expressed as

$$\pi \chi^2_{df=2;\alpha} \left(\frac{\sigma_\delta^2}{n_{K+1}} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{K} \right)^{1/2} \leq \pi \times \frac{L_t}{2} \times \frac{L_c}{2}. \tag{16}$$

Since it may be difficult to specify clinically meaningful values for L_t and L_c, it is useful to rewrite Eq. (16) as:

$$4\chi^2_{df=2;\alpha} \left(\frac{\sigma_\delta^2}{L_c^2} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{L_t^2} \right)^{1/2} \leq (n_{K+1}K)^{1/2}, \quad \text{or equivalently as}$$

$$\left(4\chi^2_{df=2;\alpha}\right)^2 \left(\frac{\sigma_\delta^2}{L_c^2} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{L_t^2} \right) \leq n_{K+1}K, \tag{17}$$

which can be rewritten further as

$$\left(4\chi^2_{df=2;\alpha}\right)^2 \left(\frac{\sigma_\delta^2}{L_c^2} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2}{L_t^2} \times \frac{\sigma_0^2 + \sigma_\varepsilon^2/n}{\sigma_0^2 + \sigma_\varepsilon^2} \right) \leq n_{K+1}K, \quad \text{or}$$

$$\left(4\chi^2_{df=2;\alpha}\right)^2 \left(\frac{1}{ES_c^2} \times \frac{1}{ES_t^2} \times \left[\frac{n-1}{n}\rho + \frac{1}{n} \right] \right) \leq n_{K+1}K. \tag{18}$$

The terms ES_c and ES_t represent the maximum allowable estimation errors for the fixed parameters of the control and treatment condition respectively, relative to the variance in each of these conditions. These are commonly used effect sizes, and what are large or small values for L_c and L_t, may be guided by a well-known classification of Cohen (1992): 0.2 is considered a small effect, 0.5 a medium effect and 0.8 a large effect.

In the treatment arm of our example, the Reconnecting Youth prevention programme is given to groups of size 9 and furthermore $\rho = 0.06$ (cf. Bauer et al., 2008). For a 95% confidence region of the fixed parameters, starting from medium effect sizes for both the control and treatment arm, that is, $ES_c = ES_t = 0.5$, by Eq. (18) it is required that $n_{K+1}K \geq 1510.47$. For the cost function in Eq. (11) the costs are minimized for a given $D_s(fixed)$ whenever $n_{K+1} = K \times \frac{c_t^*}{c_c}$ (see Eq. (D.6)). Assuming that the costs of a group in the treatment condition, c_t^*, are 10 times the cost of a person in the control condition, c_c, we obtain as values that minimize the costs: $K = 12$ groups (each of size $n = 9$) and $n_{K+1} = 126$.

The present simulation study shows that variation in group sizes leads to some efficiency loss. To restore the efficiency in a cost-efficient way, the number of persons in the control arm as well as the number of groups in the treatment arm have to be multiplied by a factor $1/RE$, where RE is the relative efficiency. Since the exact cluster size distribution is unknown, one may start from a pessimistic scenario, assuming a bimodal distribution with cluster sizes $g_a = 5$, $g_b = 10$, $g_c = 15$ and cluster frequencies $f_a = 25$, $f_b = 0$ and $f_c = 17$, for which the average group size is 9, the range of cluster sizes is 10 and the CV is 0.55. Based on the Taylor approximation the RE in terms of $D_s(fixed)$ (Eq. (7)) may become 0.96 at worst. However, according to the simulation study it is safer to lower the minimum RE by 1%, leading to a minimum RE of 0.95. The efficiency loss can thus be restored in a cost-optimal way by replicating the number of adolescents in the control arm and the number of treatment groups in the other arm $1/RE = 1/0.95 = 1.05$ times. This yields $K = 13$ and $n_{K+1} = 133$, which is a modest extension of the original design.

8. Conclusions and discussion

In analyzing the data from trials in which treatments have clustering effects, care has to be taken of the dependency between observations within clusters. For outcomes that are (approximately) normally distributed, mixed effects linear regression is a way of capturing this clustering. When comparing group therapy to pharmacological treatment or to no treatment at all, or when comparing an individual treatment condition where therapists each treat several patients to a waiting-list condition, clustering occurs in only one of two treatment arms. When planning such a trial one should consider the efficiency loss due to varying cluster sizes. Efficiency in terms of the D-criterion and in terms of the D_s-criteria for fixed effects and variance components were considered. The efficiency loss was studied by deriving expressions for the asymptotic relative efficiency of unequal versus equal cluster sizes.

To incorporate the loss of efficiency in planning a trial, second-order Taylor approximations of the asymptotic relative efficiencies were derived, of which the minima turn out to depend on the coefficient of cluster size variation only. The second-order Taylor approximations rather adequately described the asymptotic *RE*s. To the extent that the asymptotic *RE*s give an adequate description of the *RE*s for realistic sample sizes, these Taylor approximations may therefore be useful in planning trials.

In an extensive Monte Carlo simulation study, the asymptotic *RE* for fixed effects rather adequately approximated the simulated *RE*. For fixed effects, when calculating the minimum *RE* according to the Taylor approximation, it would however be more safe to lower the minimum *RE* by 1%. For the D-criterion, the minimum *RE* according to the Taylor approximation should best be lowered by 2%.

The simulated *RE* clearly depends on the coefficient of variation of the cluster sizes. The bimodal distribution and largest cluster size range yields the lowest *RE* for any of the criteria considered. The difference between ML and REML estimation was negligible for $\overline{n} = 6$ and also for $\overline{n} = 10$ in case the *RE* is defined in terms of the D_s-criterion for the fixed parameters. In other cases there was a consistent advantage of ML over REML. In these cases ML estimation thus appears to be more robust to varying cluster sizes, but note that ML estimators of the variance components are more biased (Brown and Prescott, 2006).

The simulated *RE*s show that the loss of efficiency was modest. This is to be expected, since the asymptotic relative efficiencies for the partially nested designs turned out to be larger than those for cluster randomized trials, and the minimum asymptotic relative efficiencies for cluster randomized trials have been shown to be rather high (Van Breukelen et al., 2007, 2008). For all three D-criteria the relative efficiency of unequal versus equal cluster sizes exceeds 0.94. If efficiency is defined in terms of $\text{var}(\hat{\beta}_1)$ only, the relative efficiency has been shown to become 0.90 at worst (Candel and Van Breukelen, 2009). This implies a larger efficiency loss for this criterion and thus a larger replication of the original design to regain the efficiency. However for all criteria considered, including the efficiency in terms of $\text{var}(\hat{\beta}_1)$, when planning sample sizes the (al)most cost-efficient way of regaining the efficiency loss is multiplying the number of clusters in the treatment arm as well as the number of persons in the control arm by a factor $1/RE$.

Simulations for other cluster size distributions, involving more than 3 cluster sizes and larger numbers of clusters ($K = 15$ and $K = 16$), as well as simulations involving other values for the model parameters were done. Furthermore, since ratios of the error variance in the control versus the treatment arm appear to vary between 1 and 2 (Haugli et al., 2001; Heller-Boersma et al., 2007; Roberts and Roberts, 2005), also the ratios 0.5 and 2 were examined. Since the allocation ratios for the treatment versus the control arm appear to vary between 0.4 and 1.5 in various studies (Calzone et al., 2005; Dannon et al., 2004; Haugli et al., 2001; Heller-Boersma et al., 2007; Ladouceur et al., 2000; Thompson et al., 1987; Van Minnen et al., 2003), in addition the allocation ratios 1/4 and 4 were examined. The results for these cases were in line with the results of the present Monte Carlo simulation, thereby supporting the generalizability of our conclusions.

In some intervention studies a categorical (e.g. nominal or ordinal) outcome measure is used. A useful extension of the present study would therefore involve mixed effects nominal or ordinal logistic regression. It has to be examined whether (approximate) formulas for the asymptotic relative efficiencies can be derived. Similarly to the present study, these asymptotic relative efficiencies could then be tested for their practical utility through a Monte Carlo simulation study.

Appendix A. Invariance of optimal designs under linear transformation of independent variables and ML estimation of the model parameters

Step 1: *Linear transformation of independent variables.*

The mixed effects linear regression model can be defined as (cf. Verbeke and Molenberghs, 2000) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\boldsymbol{\varepsilon}$, where $\mathbf{y}$ is an M-dimensional response vector of all subjects, $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are $(M \times P_1)$, $(M \times P_2)$ and $(M \times P_3)$ matrices of known covariates or independent variables. Further, $\boldsymbol{\beta}$ is a P_1-dimensional vector of fixed parameters, $\mathbf{u}$ is a P_2-dimensional vector of random effects and $\boldsymbol{\varepsilon}$ is a P_3-dimensional vector of residual components.

Non-degenerate linear transformations of $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ can be represented by $\mathbf{X}^* = \mathbf{X}\mathbf{Q}^{-1}$, $\mathbf{Z}_1^* = \mathbf{Z}_1\mathbf{Q}_1^{-1}$ and $\mathbf{Z}_2^* = \mathbf{Z}_2\mathbf{Q}_2^{-1}$ respectively. Note that $\mathbf{Q}$, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are square matrices that are non-singular. The mixed effects regression model can be written as:

$$\mathbf{y} = \mathbf{X}\mathbf{Q}^{-1}\mathbf{Q}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{Q}_1^{-1}\mathbf{Q}_1\mathbf{u} + \mathbf{Z}_2\mathbf{Q}_2^{-1}\mathbf{Q}_2\boldsymbol{\varepsilon} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{Z}_1^*\mathbf{u}^* + \mathbf{Z}_2^*\boldsymbol{\varepsilon}^*, \tag{A.1}$$

where $\boldsymbol{\beta}^*$, $\mathbf{u}^*$ and $\boldsymbol{\varepsilon}^*$ are the fixed effects, random effects and residual components after transformation, more precisely, $\boldsymbol{\beta}^* = \mathbf{Q}\boldsymbol{\beta}$, $\mathbf{u}^* = \mathbf{Q}_1\mathbf{u}$ and $\boldsymbol{\varepsilon}^* = \mathbf{Q}_2\boldsymbol{\varepsilon}$.

Step 2: *Covariance matrix of the ML estimators after transformation.*

We will focus on the ML estimators of the fixed effects, the proof for the variance components is similar. Since ML estimators are invariant under linear transformation (Mood et al., 1974) and premultiplication by $\mathbf{Q}$ is a linear transformation, we have for the ML estimators of the fixed effects before, $\hat{\boldsymbol{\beta}}$, and after the transformation, $\hat{\boldsymbol{\beta}}^*$, $\hat{\boldsymbol{\beta}}^* = \mathbf{Q}\hat{\boldsymbol{\beta}}$ and thus $\text{Var}\,\hat{\boldsymbol{\beta}}^* = \mathbf{Q}\text{Var}\,\hat{\boldsymbol{\beta}}\mathbf{Q}^T$. As an example, to transform the model in Eq. (1) into the model in Eq. (12), we have $\mathbf{Q}^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$ and thus $\mathbf{Q} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. For the ML estimators of the fixed parameters in Eq. (12), $\hat{\boldsymbol{\beta}}^*$, we then obtain:

$$\text{Var}\,\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \text{Var}\,\hat{\boldsymbol{\beta}} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{A.2}$$

Step 3: *Determinant of ML estimators of the fixed effects after transformation.*

Let $\hat{\boldsymbol{\theta}}_f^*$ denote the ML estimators of the fixed effect estimators after transformation (i.e. $\hat{\boldsymbol{\beta}}^*$). Furthermore, let $|\mathbf{A}|$ denote the determinant of matrix $\mathbf{A}$ and let $T(\xi)$ denote the design ξ after transformation of the matrices $\mathbf{X}$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$. We then have (see e.g. Harville, 1997):

$$\text{Det}(\text{Var}(\hat{\boldsymbol{\theta}}_f^* \mid T(\xi))) = \left|\text{Var}(\hat{\boldsymbol{\beta}}^* \mid T(\xi))\right| = |\mathbf{Q}| \left|\text{Var}(\hat{\boldsymbol{\beta}} \mid \xi)\right| \left|\mathbf{Q}^T\right|. \tag{A.3}$$

Step 4: *Relative efficiency of two designs for the fixed effects after transformation.*

If we consider the relative efficiency $\left(\frac{\text{Det}(\text{Var}(\hat{\boldsymbol{\theta}}_f^* \mid T(\xi_1)))}{\text{Det}(\text{Var}(\hat{\boldsymbol{\theta}}_f^* \mid T(\xi_2)))}\right)$ for two different designs ξ_1 and ξ_2 after linear transformation, then, making use of the result in Eq. (A.3), we obtain:

$$\frac{|\mathbf{Q}| \left|\text{Var}(\hat{\boldsymbol{\beta}} \mid \xi_1)\right| \left|\mathbf{Q}^T\right|}{|\mathbf{Q}| \left|\text{Var}(\hat{\boldsymbol{\beta}} \mid \xi_2)\right| \left|\mathbf{Q}^T\right|} = \frac{\left|\text{Var}(\hat{\boldsymbol{\beta}} \mid \xi_1)\right|}{\left|\text{Var}(\hat{\boldsymbol{\beta}} \mid \xi_2)\right|}, \tag{A.4}$$

which is the same as the relative efficiency $\left(\frac{\text{Det}(\text{Var}(\hat{\boldsymbol{\theta}}_f \mid \xi_1))}{\text{Det}(\text{Var}(\hat{\boldsymbol{\theta}}_f \mid \xi_2))}\right)$ before transformation. So, the order of designs in terms of $D_s(fixed)$, and thus also the optimal design does not change after transformation.

A similar proof can be given to show that the optimality of a design in terms of $D_s(random)$ does not change after linear transformation. Finally, note that, since the ML estimators of fixed effects and variance components asymptotically are independent (McCulloch and Searle, 2001), the determinant of all parameter estimates factorizes into $D_s(fixed)$ and $D_s(random)$. The ratio of the D-criterion for two different designs is thus the product of two efficiency ratios in terms of $D_s(fixed)$ and $D_s(random)$, and therefore will also not change after linear transformation. Hence, the optimality of a design in terms of the D-criterion also is invariant under linear transformations of the independent variables.

Appendix B. Derivation of the RE for ML estimators of the fixed parameters

Let the mixed effects linear regression model be defined by Eq. (1). Furthermore, define $w_j = \frac{n_j}{n_j\sigma_0^2+\sigma_\varepsilon^2}$. As shown by Candel and Van Breukelen (2009), the variance–covariance matrix for the ML-estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by:

$$\text{Var}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{\sigma_\delta^2}{n_{K+1}} \begin{bmatrix} 1 & -1 \\ -1 & 1 + \dfrac{n_{K+1}}{\sigma_\delta^2 \sum_{j=1}^{K} w_j} \end{bmatrix}. \tag{B.1}$$

The determinant of the Var matrix is:

$$\text{Det}\left(\text{Var}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right) = \frac{\sigma_\delta^2}{n_{K+1} \sum_{j=1}^{K} w_j}. \tag{B.2}$$

Applying the transformation in Eq. (A.2) and employing Eq. (B.1), we obtain the variance–covariance matrix for the ML-estimators $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ in Eq. (12) as:

$$\text{Var}\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \frac{\sigma_\delta^2}{n_{K+1}} \begin{bmatrix} 1 & -1 \\ -1 & 1 + \dfrac{n_{K+1}}{\sigma_\delta^2 \sum_{j=1}^{K} w_j} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \frac{\sigma_\delta^2}{n_{K+1}} \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{n_{K+1}}{\sigma_\delta^2 \sum_{j=1}^{K} w_j} \end{bmatrix}. \tag{B.3}$$

The determinant of the variance–covariance matrix in Eq. (B.3) also is given by Eq. (B.2).

Let $w_e = \frac{\bar{n}}{\bar{n}\sigma_0^2+\sigma_\varepsilon^2}$, with $\bar{n}$ the average cluster size of the clusters in the treatment condition. The relative efficiency as defined by Eq. (3), keeping N and n_{K+1} constant, can now be established, both for the model in Eqs. (1) and (12), as:

$$RE(D_s(fixed)) = \sqrt{\frac{\sum_{j=1}^{K} w_j}{K w_e}} = \sqrt{\frac{\bar{n}+(1-\rho)/\rho}{\bar{n}} \times \frac{1}{K}\sum_{j=1}^{K}\left(\frac{n_j}{n_j+(1-\rho)/\rho}\right)}. \tag{B.4}$$

Employing the Jensen inequality (see e.g. Mood et al., 1974, p. 72), it can be seen that Eq. (B.4) does not exceed 1. This implies that unequal cluster sizes are less efficient than equal cluster sizes for the fixed effect estimators. Also, the relative efficiency in Eq. (B.4) is larger than the corresponding relative efficiency for cluster randomized trials (see Van Breukelen et al., 2007).

Appendix C. Derivation of the RE for ML estimators of the variance components

The variance components of the treatment arm and the control arm, are estimated from separate parts of the sample, and therefore are independent. Let $\mathbf{0}^T = [0, 0]$, we then have:

$$D_s(random) = \mathrm{Det}\left(\mathrm{Var}\begin{bmatrix}\hat{\sigma}_0^2\\ \hat{\sigma}_\varepsilon^2\\ \hat{\sigma}_\delta^2\end{bmatrix}\right) = \mathrm{Det}\begin{bmatrix}\mathrm{Var}\begin{bmatrix}\hat{\sigma}_0^2\\ \hat{\sigma}_\varepsilon^2\end{bmatrix} & \mathbf{0}\\ \mathbf{0}^T & \mathrm{Var}(\hat{\sigma}_\delta^2)\end{bmatrix} = \mathrm{Det}\left(\mathrm{Var}\begin{bmatrix}\hat{\sigma}_0^2\\ \hat{\sigma}_\varepsilon^2\end{bmatrix}\right) \times \mathrm{Var}(\hat{\sigma}_\delta^2). \tag{C.1}$$

The determinant of the asymptotic variance–covariance matrix of ML estimators of the variance components for the treatment condition is given by (see Candel et al., 2008, p. 236):

$$\mathrm{Det}\left(\mathrm{Var}\begin{bmatrix}\hat{\sigma}_0^2\\ \hat{\sigma}_\varepsilon^2\end{bmatrix}\right) = \frac{4\sigma_\varepsilon^2}{N\sum_{j=1}^{K} w_j^2 - \left(\sum_{j=1}^{K} w_j\right)^2}. \tag{C.2}$$

The ML estimator of the error variance in the control condition without clusters, that is, σ_δ^2, is given by (see e.g. Mood et al., 1974, p. 281):

$$\hat{\sigma}_\delta^2 = \frac{\sum_{i=1}^{n_{K+1}}(y_{iK+1}-\bar{y}_{K+1})^2}{n_{K+1}}, \quad \text{with } \bar{y}_{K+1} = \frac{\sum_{i=1}^{n_{K+1}} y_{iK+1}}{n_{K+1}}. \tag{C.3}$$

It is known that, if y_{iK+1} $(i = 1, \ldots, n_{K+1})$ is independently normally distributed with the same mean and variance σ_δ^2, then $\hat{\sigma}_\delta^2 \times \frac{n_{K+1}}{\sigma_\delta^2}$ is chi-square distributed, with degrees of freedom equal to $n_{K+1}-1$ (Mood et al., 1974, p. 245). Since the variance of such a chi-square distributed variable is equal to $2(n_{K+1}-1)$, the variance of $\hat{\sigma}_\delta^2$ is equal to:

$$\mathrm{Var}\left(\hat{\sigma}_\delta^2\right) = \frac{2(n_{K+1}-1)\sigma_\delta^4}{n_{K+1}^2}, \quad \text{which for } n_{K+1}\to\infty \text{ equals } \frac{2\sigma_\delta^4}{n_{K+1}}. \tag{C.4}$$

This implies that asymptotically (see Eq. (C.1)):

$$D_s(random) = \left(\frac{4\sigma_\varepsilon^4}{N\sum_{j=1}^{K} w_j^2 - \left(\sum_{j=1}^{K} w_j\right)^2}\right) \times \frac{2\sigma_\delta^4}{n_{K+1}}. \tag{C.5}$$

For the relative efficiency, keeping N and n_{K+1} constant, we then have

$$RE\left(D_s(random)\right) = \left(\frac{N\sum_{j=1}^{K} w_j^2 - \left(\sum_{j=1}^{K} w_j\right)^2}{(N-K)Kw_e^2}\right)^{1/3}, \tag{C.6}$$

which, if taken to the power 3/2, is the same as $RE(D_s(random))$ for cluster randomized trials (see Van Breukelen et al., 2008).

Appendix D. Derivation of cost-efficient replication to regain the efficiency loss due to varying cluster sizes

Let f_1 and f_2 denote the replication factors of the number of persons in the control arm and the number of groups in the treatment arm, respectively. We will first consider $D_s(fixed)$. Let RE denote the relative efficiency of unequal versus equal group sizes for this criterion. As can be seen in Eq. (B.2), to restore the relative efficiency in terms of $D_s(fixed)$, the replication factors should satisfy $f_1 \times f_2 = 1/RE^2$, or $f_2 = 1/(f_1 \times RE^2)$. We have to choose f_1 and f_2 such that the cost function, as defined in Eq. (11), that is

$$f_1 n_{K+1} c_c + f_2 K c_t^* = f_1 n_{K+1} c_c + K c_t^* / (f_1 \times RE^2), \tag{D.1}$$

is minimized. Setting the first derivative with respect to f_1 equal to 0 yields:

$$n_{K+1} c_c - K c_t^* / (f_1^2 \times RE^2) = 0, \quad \text{or} \quad f_1^2 = \frac{K c_t^*}{n_{K+1} c_c} \frac{1}{RE^2}. \tag{D.2}$$

In case of equal group sizes with $n_j = \bar{n}$ for all $j = 1, \ldots, K$, one can rewrite Eq. (B.2) as:

$$D_s(fixed) = \sigma_\delta^2 / (n_{K+1} K w_e), \tag{D.3}$$

and the cost function (see Eq. (11)) can then be rewritten as

$$C = \frac{c_c}{D_s(fixed)} \times \frac{\sigma_\delta^2}{K w_e} + K c_t^*. \tag{D.4}$$

Setting the derivative of Eq. (D.4) with respect to K equal to 0 yields

$$-\frac{c_c}{D_s(fixed)} \times \frac{\sigma_\delta^2}{K^2 w_e} + c_t^* = 0, \tag{D.5}$$

or equivalently, by substituting Eq. (D.3) for $D_s(fixed)$, $c_t^*/c_c = n_{K+1}/K$ which yields a minimum for the costs C. To summarize, when planning a design with equal group sizes so as to minimize the costs for a given value of $D_s(fixed)$, the ratio of the number of persons in the control arm versus the number of groups in the treatment arm will satisfy:

$$\frac{n_{K+1}}{K} = \frac{c_t^*}{c_c}. \tag{D.6}$$

Combining Eqs. (D.2) and (D.6) yields the optimal replication factors to restore the efficiency loss for such a design: $f_1 = 1/RE$, and thus $f_2 = 1/(f_1 \times RE^2) = 1/RE$.

To restore the relative efficiency in terms of $D_s(random)$ and the D-criterion, the replication factors should satisfy $f_1 \times f_2^2 = 1/RE^3$ (see Eq. (C.5)) and $f_1^2 \times f_2^3 = 1/RE^5$, with RE denoting the relative efficiencies for $D_s(random)$ and the D-criterion respectively. By similar derivations as given for $D_s(fixed)$, we can show that these efficiencies are restored at the lowest costs, again by choosing $f_1 = f_2 = 1/RE$.

References

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Design. Oxford University Press, Oxford.

Bauer, D.J., Sterba, S.K., Hallfors, D.D., 2008. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behavioral Research 43, 210–236.

Brown, H., Prescott, R., 2006. Applied Mixed Models in Medicine. Wiley, Chichester.

Calzone, K.A., Prindiville, S.A., Jourkiv, O., Jenkins, J., DeCarvalho, M., Wallerstedt, D.B., Leiwehr, D.J., Steinberg, S.M., Soballe, P.W., Lipkowitz, S., Klein, P., Kirsch, I.R., 2005. Randomized comparison of group versus individual genetic education and counseling for familial breast and/or ovarian cancer. Journal of Clinical Oncology 23, 3455–3464.

Candel, M.J.J.M., Van Breukelen, G.J.P., 2009. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. Statistics in Medicine 28, 2307–2324.

Candel, M.J.J.M., Van Breukelen, G.J.P., Kotova, L., Berger, M.P.F., 2008. Optimality of unequal cluster sizes in multilevel studies with small sample sizes. Communications in Statistics: Simulation and Computation 37, 222–239.

Cohen, J., 1992. A power primer. Psychological Bulletin 112, 155–159.

Dannon, P.N., Gon-Usishkin, M., Gelbert, A., Lowengrub, K., Grunhaus, L., 2004. Cognitive behavioral group therapy in panic disorder patients: the efficacy of CBGT versus drug treatment. Annals of Clinical Psychiatry 16, 41–46.

Demidenko, E., 2004. Mixed Models: Theory and Applications. Wiley, Hoboken, NJ.

Donner, A., Klar, N., 1994. Cluster randomization trials in epidemiology: theory and application. Journal of Statistical Planning and Inference 42, 37–56.

Harville, D.A., 1997. Matrix Algebra from a Statistician's Perspective. Springer, New York.

Haugli, L., Steen, E., Laerum, E., Nygard, R., Finset, A., 2001. Learning to have less pain—Is it possible? A one-year follow-up study of the effects of a personal construct group learning programme on patients with chronic musculoskeletal pain. Patient Education and Counseling 45, 111–118.

Heller-Boersma, J.G., Schmidt, U.H., Edmonds, D.K., 2007. A randomized controlled trial of a cognitive-behavioural group intervention versus waiting-list control for women with uterovaginal agenesis (Mayer–Rokitansky–Küster–Hauser syndrome: MRKH). Human Reproduction 22, 2296–2301.

Jarrett, R.B., Schaffer, M., McIntire, D., Witt-Browder, A., Kraft, D., Risser, R.C., 1999. Treatment of atypical depression with cognitive therapy or phenelzine. A double-blind, placebo-controlled trial. Archives of General Psychiatry 56, 431–437.

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Analysis. Prentice Hall, Upper Saddle River, NJ.

Kessels, R., Goos, P., Vandebroek, M., 2008. Optimal designs for conjoint experiments. Computational Statistics and Data Analysis 52, 2369–2387.

Ladouceur, R., Dugas, M.J., Freeston, M.H., Léger, E., Gagnon, F., Thibodeau, N., 2000. Efficacy of a cognitive-behavioral treatment for generalized anxiety disorder: Evaluation in a controlled trial. Journal of Consulting and Clinical Psychology 68, 957–964.
McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. Wiley, New York.
Moerbeek, M., 2005. Robustness properties of *A*-, *D*-, and *E*-optimal designs for polynomial growth models with autocorrelated errors. Computational Statistics and Data Analysis 48, 765–778.
Moerbeek, M., Wong, W.K., 2008. Sample size formulae for trials comparing group and individual treatments in a multilevel model. Statistics in Medicine 27, 2850–2864.
Mood, A.M., Graybill, F.A., Boes, D.C., 1974. Introduction to the Theory of Statistics, third ed. McGraw-Hill, Tokyo.
Ortega-Azurduy, S.A., Tan, F.E.S., Berger, M.P.F., 2008. The effect of dropout on the efficiency of *D*-optimal designs of linear mixed models. Statistics in Medicine 27, 2601–2617.
Pals, S.L., 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.
Parker, D.R., Evangelou, E., Eaton, C.B., 2005. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). Contemporary Clinical Trials 26, 260–267.
Pisinger, C., Vestbø, J., Borch-Johnsen, K., Jørgensen, T., 2005. Smoking cessation intervention in a large randomised population-based study. The Inter99 study. Preventive Medicine 40, 285–292.
Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T., 2000. A user's guide to MlwiN, second ed., Centre for Multilevel Modelling, Institute of Education, London.
Raudenbush, S.W., 1997. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods 2, 173–185.
Roberts, C., 1999. The implication of variation in outcome between health professionals for the design and analysis of randomized controlled trials. Statistics in Medicine 18, 2605–2615.
Roberts, C., Roberts, S.A., 2005. The design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2, 152–162.
Rosner, B., 2006. Fundamentals of Biostatistics. Thomson, Belmont, CA.
Smeeth, L., Siu-Woon, E., 2002. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC trial of the assessment and management of older people in the community. Controlled Clinical Trials 23, 409–421.
Tekle, F.B., Tan, F.E.S., Berger, M.P.F., 2008. Maximin *D*-optimal designs for binary longitudinal responses. Computational Statistics and Data Analysis 52, 5253–5262.
Thompson, L.W., Gallagher, D., Breckenridge, J.S., 1987. Comparative effectiveness of psychotherapies for depressed elders. Journal of Consulting and Clinical Psychology 55, 385–390.
Van Berkestein, L.G.M., Kastein, M.R., Lodder, A., DeMelker, R.A., Bartelink, M.L., 1999. How do we compare with our colleagues? Quality of general practitioner performance in consultations for non-acute abdominal complaints. International Journal for Quality in Health Care 11, 475–486.
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.
Van Breukelen, G.J.P., Candel, M.J.J.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.
Van Minnen, A., Hoogduin, K.A.L., Keijsers, G.P.J., Hellenbrand, I., Hendriks, G.-J., 2003. Treatment of trichotillomania with behavioral therapy or fluoxetine. Archives of General Psychiatry 60, 517–522.
Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.
Wampold, B.E., Serlin, R.C., 2000. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. Psychological Methods 5, 425–433.