UNDERSTANDING STATISTICS, *2*(3), 151–175
Copyright © 2003, Lawrence Erlbaum Associates, Inc.

Optimal Sample Sizes in Experimental Designs With Individuals Nested Within Clusters

Mirjam Moerbeek
Department of Methodology and Statistics
Utrecht University

Gerard J. P. Van Breukelen and Martijn P. F. Berger
Department of Methodology and Statistics
Maastricht University

Marlein Ausems
Department of Health Education and Promotion
Maastricht University

This article deals with optimal sample sizes for experimental studies with individuals nested within clusters. Examples are multicenter clinical trials with patients nested within clinics or family practices, studies on absence due to illness with employees nested within work sites, or school-based smoking prevention interventions with pupils nested within schools. Optimal sample sizes are given for models with continuous or binary outcomes and for 1 or 2 treatment factors. Because individuals are nested within clusters, not only the total number of individuals needed to obtain a specified power of the test of treatment effect has to be established but also the number of clusters and the number of individuals per cluster. We show that different outcome variables lead to different optimal sample sizes. Furthermore, the statistically optimal design is not always feasible in practice because of certain limitations, and we show how to calculate the optimal design given restrictions on the sample sizes at both the cluster and individual level. Also, the dropout of clusters and individuals are taken into account.

Requests for reprints should be sent to Mirjam Moerbeek, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: M.Moerbeek@FSS.UU.NL

Keywords: optimal sample sizes, standard errors, power,
smoking prevention intervention

In many experimental studies, individuals are nested within clusters. Examples are multicenter clinical trials with patients nested within clinics or family practices, drug use cessation intervention studies with clients nested within therapy groups, studies on absence due to illness with employees nested within work sites, or smoking prevention intervention studies with pupils nested within schools.

An experimental study aims at estimating and detecting an existing treatment effect with high precision and power, respectively. There are several strategies to increase statistical power; the most obvious one is increasing the number of observations at the cluster and individual level. This strategy, however, is limited by the study budget because costs are associated with including individuals and clusters in the study. In addition to the study costs, the time and effort of the participants that are willing to participate must be used efficiently. In the case of individuals nested within clusters, the design is determined by the allocation of units (i.e., the sample size at both the cluster and individual level). The optimal design for an experimental evaluation of an intervention will often be the design that leads to the smallest standard error of the estimated treatment effect and consequently to the largest statistical power of the test on treatment effect and narrowest confidence interval for the treatment effect, assuming unbiased estimation of this effect, of course. Using an optimal design is important because nonoptimal designs lead to two situations that are increasingly unacceptable (for a review of the basic concepts of optimal design, see McClelland, 1997). First, the number of participants is increased to compensate for design inefficiencies when it should be minimized for both ethical and practical reasons. Second, the power to detect treatment effects is reduced.

The aim of this article was to show how optimal designs for the case of individuals nested within clusters may be obtained and how nonoptimal designs lead to a waste of money and a misuse of the willingness of participants to participate in the study. Sample size formulae to achieve a certain power of the test on treatment effect are available for continuous outcome variables under the assumption that either the number of clusters or the number of individuals per cluster is fixed (Donner, Bircket, & Buck, 1981; Hsieh, 1988; Siddiqui, Hedeker, Flay, & Hu, 1996). It is, however, also possible and sometimes more realistic to calculate optimal sample sizes at both levels (clusters and individuals) given a certain budget for sampling and measuring individuals and clusters, as has been done for continuous outcome variables by Raudenbush (1997) and Snijders and Bosker (1993). This budget may be calculated such that a certain statistical power is achieved for the test of the intervention treatment effect.

In this article, we make a distinction between three steps in planning an experimental study for the case of individuals nested within clusters. The first step is the

selection of the statistical model for the data analysis. The multilevel model, which is given in the next section, may be used for nested data because it accounts for the nesting of individuals within clusters and the dependency of the outcomes of individuals within the same cluster. The linear multilevel model can be used for continuous outcomes, whereas binary outcome variables can be analyzed with the multilevel logistic model. Relevant covariates must be selected beforehand and included into the model as well.

The costs for the enrolment of clusters and individuals, the dropout rates at the individual and cluster level, as well as the values of the parameters in the statistical model that will be used to analyze the data have to be known in advance to calculate the optimal number of schools and pupils per school. This means that intervention studies are developed and implemented to get some knowledge of the parameter values, notably the intervention effect, but these values have to be known in advance to plan the study as optimally as possible. To solve this problem, a reasonable prior estimate of these parameter values may be used instead of the true parameter values. Such an estimate may be obtained from a pilot study, from results of comparable intervention studies, and from theoretical opinions about the minimally relevant treatment effect. This is the second step in planning an intervention study.

The third step is to calculate the optimal sample sizes at both levels and the corresponding standard error, which in turn determines the power of the test of no treatment effect. The optimal sample sizes may be restricted by the actual number of schools and number of pupils per school and these restrictions may be used as preconditions when calculating the sample sizes. Furthermore, the standard error will increase due to dropout at both the pupil and school level; therefore, dropout rates should be taken into account when planning the experimental design and budget. It also has to be noted that different outcome variables may lead to different optimal designs, so one should establish the optimal design for the outcome variable of most importance.

We illustrate the process of planning an intervention study using data from a Dutch school-based smoking prevention intervention (Ausems, Mesters, Van Breukelen, & De Vries, 2002). In the next section, the multilevel model is given and the data are analyzed to get some knowledge of the values of the parameters. These parameter values are needed to calculate the optimal sample sizes at the cluster and individual level, as is shown in the Optimal Sample Sizes section. This section also shows how the power of detecting a treatment effect can be calculated. The results in this section are based on the papers on optimal designs mentioned previously and on the papers by Moerbeek, Van Breukelen, and Berger (2000, 2001a, 2001b). In the second to last section, optimal designs for the Dutch smoking prevention intervention were derived using the parameter estimates from the Multilevel Model section, and we show how nonoptimal designs result in a loss of statistical power.

MULTILEVEL MODEL

Description of the Intervention Study Used as Example

In this section, we introduce and explain multilevel modelling of nested data by using as example the Dutch school-based smoking prevention intervention study by Ausems et al. (2002). This study aimed at preventing and delaying the onset of smoking and therefore targeted elementary school pupils aged 11 and 12 years (the highest group of the Dutch elementary schools) because research indicated that the smoking Dutch youth had smoked their first cigarette at that age. The study evaluated the effects of two interventions: (a) an existing in-school prevention program in which pupils received information on different aspects of smoking (e.g., smoking and health, costs of smoking, advertisements) during lessons and (b) an out-of-school intervention in which pupils received personalized tailored letters on smoking. These interventions were evaluated simultaneously because research indicated that the in-school intervention would function more effectively when it was supplemented with an out-of-school intervention. Therefore, there were two treatment factors with two levels each, which resulted in a total of four treatment groups: a group receiving the in-school intervention (34 schools), a group receiving the out-of-school intervention (36 schools), a group receiving both interventions (36 schools), and a control group (34 schools). Randomization to treatment groups was done at the school level. A total of 3,349 pupils participated in the study; the number of pupils per school ranged from 3 to 53, with a mean of 23.9, and 79% of the sample sizes at the school level ranged between 15 and 35. We used two outcome measures at the first posttest to illustrate the results on optimal designs in our article, namely the continuous variable "attitude toward the disadvantages of smoking," and the binary variable "smoking behavior."

If there were no nesting of pupils within schools, this study could have been analyzed with two-way analysis of variance (ANOVA) or linear regression for continuous outcomes and with logistic regression for binary outcomes. However, pupils were nested within schools and it is reasonable to assume that smoking behavior of pupils within the same school is correlated due to mutual influence and school policy toward smoking and similarly for other outcomes such as attitude. Traditional regression analysis, which ignores nesting of pupils within schools and dependency of outcomes within a school, will incorrectly estimate standard errors of predictor variables, and this may lead to incorrect conclusions on the effect of an intervention. The data analysis should take into account the nesting of pupils within schools and the dependency of outcomes within the same school, and this may be done using the multilevel modeling in which school effects on the outcome variable are included as random effect similar to the random error or residual at the pupil level in ordinary linear regression (Goldstein, 1995; Hox, 2002; Kreft & De Leeuw, 1998; Longford, 1995; Raudenbush & Bryk, 2002; Snijders & Bosker,

1999). Multilevel models may be used when the schools in the intervention study can be looked on as a random sample from a population of schools. For a review of multilevel and more traditional models for nested data, we refer to Moerbeek, Van Breukelen, and Berger (2003a).

Multilevel Linear Regression of a Quantitative Outcome

The basic idea of multilevel modeling is that regression models are formulated at each level of the multilevel data structure and that the models may be combined into one single model. For instance, consider the continuous variable attitude toward the disadvantages of smoking (ATTITUDE), which is the summation of 11 items measured on a 5-point scale ranging from 1 (*positive*) to 5 (*negative*) toward the disadvantages of smoking. Thus, the sum score ranges from 11 to 55. The pupil level model relates the outcome variable ATTITUDE to relevant covariates at that level: PRE_ATTITUDE (attitude toward the disadvantages of smoking at pretest), AGE, and GENDER. The model for pupil i within school j is then

$$\begin{aligned} \text{ATTITUDE}_{ij} = {} & \beta_{0j} + \beta_{1j}\text{PRE\_ATTITUDE}_{ij} + \\ & \beta_{2j}\text{AGE}_{ij} + \beta_{3j}\text{GENDER}_{ij} + e_{ij}\,. \end{aligned} \tag{1}$$

Note that this pupil-level model describes outcome variation within school j as a function of pupil characteristics and does not include the two treatment factors because schools were randomized and therefore treatment does not vary within schools. Instead, treatment appears in the school-level part of the model later on. The random error terms e_{ij} are independently and normally distributed with zero mean and variance σ_e^2. The pretest measurement on attitude was included into the model to correct for pretest differences in attitude, thereby reducing unexplained variance and increasing statistical power. Age and gender were included because it was thought that these are related to the attitude toward smoking. Multilevel models differ from traditional regression models in that they allow each school to have its own intercept β_{0j}, which means that the mean level of the outcome variable differs across schools. Moreover, each school may also have its own slopes: β_{1j}, β_{2j}, and β_{3j}. In that case, we have an interaction of predictor variable by cluster, that is, the effect of the predictor variable varies between clusters. Both the intercept and slopes may vary as a function of school-level covariates (e.g. high schools vs. vocational schools, urban vs. rural schools, treated vs. untreated schools) and a random school effect. We assumed that the slopes do not vary (which was confirmed by a preliminary data analysis) and that the intercept varies as a function of the treatment conditions and a random school effect. The school level model is then

$$\begin{aligned}\beta_{0j} &= \beta_0 + \beta_4 \mathrm{IN}_j + \beta_5 \mathrm{OUT}_j + \beta_6 \mathrm{INTERAC}_j + u_{0j} \\ \beta_{1j} &= \beta_1 \\ \beta_{2j} &= \beta_2 \\ \beta_{3j} &= \beta_3 ,\end{aligned} \tag{2}$$

where IN, OUT, and INTERAC represent treatment main effects and interaction, respectively (see following for details). The first equation in Equation 2 predicts the intercept (i.e., the overall level of the outcome variable) from the experimental conditions of the design. The next equation states that the overall slope β_{1j} for the effect of the pretest measurement on attitude is constant across all schools, that is, no pretest by school interaction is assumed. The third and fourth equation of Equation 2 state that the same holds for the effect of age on posttest attitude and for the effect of gender on posttest attitude, respectively. The random term u_{0j} is the random (unexplained) deviation of the average outcome in school j from the overall average outcome of all schools with the same values of all covariates at school and pupil level and reflects unexplained differences in posttest attitude between schools within the same treatment condition adjusted for pupil age, gender, and pretest attitude. It is therefore similar to the random term e_{ij} at the pupil level. The u_{0j} are independently and normally distributed with zero mean and variances σ^2_{u0}. Moreover, the u_{0j} are also independent of e_{ij}.

The variable IN = +0.5 for a school participating in the in-school intervention and –0.5 otherwise so that this variable is centered around zero and β_4 measures the effect of the in-school intervention averaged across both levels of the OUT factor. The variable OUT for the out-of-school intervention is coded likewise, and the interaction term INTERAC is calculated as INTERAC = 2*IN*OUT and therefore has values –0.5 and +0.5 as well. This coding scheme has some advantages compared with (0, 1) coding, namely that the IN and OUT predictors are uncorrelated with the interaction term, and the regression coefficients β_4, β_5, and β_6 reflect orthogonal contrasts between pairs of treatment conditions similar to two-way ANOVA (see following for details) and that relatively simple formulae for the estimators of the regression coefficients and their standard errors are obtained (Moerbeek et al., 2001b). Substitution of the school-level model into the pupil-level model gives the single equation model

$$\begin{aligned}\mathrm{ATTITUDE}_{ij} = \beta_0 &+ \beta_1 \mathrm{PRE\_ATTITUDE}_{ij} + \beta_2 \mathrm{AGE}_{ij} + \beta_3 \mathrm{GENDER}_{ij} \\ &+ \beta_4 \mathrm{IN}_j + \beta_5 \mathrm{OUT}_j + \beta_6 \mathrm{INTERAC}_j + u_{0j} + e_{ij} .\end{aligned} \tag{3}$$

Table 1 shows the coding scheme and the average outcomes for the four treatment groups. As follows from Table 1, β_4 is equal to the difference in average outcomes in Groups 3 and 4 on one hand and Groups 1 and 2 on the other. Likewise,

TABLE 1
Coding Scheme and Average Outcomes for the Four Intervention Groups

Group	*Intervention*	*IN*	*OUT*	*INTERAC*	*Average Outcome*
1	None	–0.5	–0.5	+0.5	$\beta_0 + \beta_1$ PRE_ATTITUDE + β_2AGE + β_3GENDER – $0.5\beta_4$ – $0.5\beta_5$ – $0.5\beta_6$
2	Out-of-school	–0.5	+0.5	–0.5	$\beta_0 + \beta_1$ PRE_ATTITUDE + β_2AGE + β_3GENDER – $0.5\beta_4$ + $0.5\beta_5$ – $0.5\beta_6$
3	In-school	+0.5	–0.5	–0.5	$\beta_0 + \beta_1$ PRE_ATTITUDE + β_2AGE + β_3GENDER + $0.5\beta_4$ – $0.5\beta_5$ – $0.5\beta_6$
4	Both	+0.5	+0.5	+0.5	$\beta_0 + \beta_1$ PRE_ATTITUDE + β_2AGE + β_3GENDER + $0.5\beta_4$ + $0.5\beta_5$ + $0.5\beta_6$

Note. The variable INTERAC is calculated as INTERAC = 2*IN*OUT.

β_5 is equal to the difference in average outcomes in Groups 2 and 4 on one hand and Groups 1 and 3 on the other, and β_6 is equal to the difference in average outcomes in Groups 1 and 4 on one hand and Groups 2 and 3 on the other. Therefore, β_4, β_5, and β_6 reflect average differences between pairs of treatment conditions and come down to the main effects of both interventions and their interaction, well known in traditional two-way ANOVA. The residual or unexplained variance of the outcomes in any treatment group is equal to $\sigma^2_{u0} + \sigma^2_e$. The variances σ^2_{u0} and σ^2_e are called *variance components* because they both contribute to the total variance. The residual covariance of outcomes within the same school is equal to σ^2_{u0}. The intraschool correlation coefficient ρ measures the relative amount of variance at the school level and is calculated as $\rho = \sigma^2_{u0} / (\sigma^2_{u0} + \sigma^2_e)$. In general, this is referred to as the *intraclass correlation.*

The computer program MIXREG (Hedeker & Gibbons, 1996b) was used for the data analysis, but we could also have used SAS PROC MIXED (Little, Milliken, Stroup, & Wolfinger, 1996) or the specialized computer packages for multilevel analysis MLwiN (Goldstein et al., 2000) or HLM (Bryk, Raudenbush, & Congdon, 1996). The variable GENDER was found not to have a significant effect on ATTITUDE and was excluded from the model. The results of the analysis are given in Table 2. Note that the intraschool correlation coefficient was equal to $\rho = \sigma^2_{u0} / (\sigma^2_{u0} + \sigma^2_e) = 3.349 / (3.349 + 44.952) = 0.069$.

Multilevel Logistic Regression of a Binary Outcome

The multilevel model for the binary outcome variable SMOKE assumes that

TABLE 2
Predictors of Attitude Toward Disadvantages of Smoking

Variable	*Estimate*	*Standard Error*	*Z*	*p Value*
Intercept	$\hat{\beta}_0 = 24.198$	2.647	9.143	.000
Pretreatment attitude	$\hat{\beta}_1 = 0.564$	0.016	35.343	.000
Age	$\hat{\beta}_2 = -0.610$	0.217	–2.809	.005
Condition				
Program (IN)	$\hat{\beta}_4 = 1.566$	0.395	3.963	.000
Tailor (OUT)	$\hat{\beta}_5 = 2.395$	0.395	6.070	.000
Interaction (INTERAC)	$\hat{\beta}_6 = -1.259$	0.395	–3.191	.001
Random school effect	$\hat{\sigma}^2_{u0} = 3.349$	0.642		
Random pupil effect	$\hat{\sigma}^2_e = 44.952$	1.122		

Note. The variable IN is coded +0.5 for a pupil receiving the in-school intervention and –0.5 otherwise. The variable OUT is coded likewise, and the variable is calculated as INTERAC = 2*IN*OUT, therefore, it has values –0.5 and +0.5 as well.

$$\log\left(\frac{\pi_{ij}}{1- \pi_{ij}}\right) = \beta_0 + \beta_1 \text{PRE\_SMOKE}_{ij} + \beta_2 \text{AGE}_{ij} + \beta_3 \text{GENDER}_{ij} + \beta_4 \text{IN}_j + \beta_5 \text{OUT}_j + \beta_6 \text{INTERAC}_j + u_{0j}, \quad (4)$$

where π_{ij} is the probability of being a smoker. PRE_SMOKE is the smoking behavior at pretest.

Estimation methods for the multilevel logistic model are not yet as well developed as those for the linear multilevel model. The computer program MLwiN (Goldstein et al., 2000) offers the approximate methods Marginal Quasi Likelihood estimator and Penalized Quasi Likelihood. Rodríguez and Goldman (2001) concluded that second order Penalized Quasi Likelihood estimator provides the best approximation but that the estimates of the fixed and random effects may still be underestimated when the random effects are sizeable. Moerbeek, Van Breukelen, and Berger (2003b) performed an extensive simulation study and showed that the bias of the estimated treatment effect and its standard error were in general smaller than 5% and 7%, respectively, when using second order Penalized Quasi Likelihood. Also, no convergence problems occurred when using this method as opposed to estimation by numerical integration by using MIXOR (Hedeker & Gibbons, 1996a). We therefore decided to use second order Penalized Quasi Likelihood in this article; the results of the data analysis can be found in Table 3.

Estimating Equation 4 with second order Penalized Quasi Likelihood showed no interaction or gender effect, which were therefore dropped from the model. However, the reduced model with IN, OUT, AGE, and pretest smoking as predictors is still too complex for existing optimal sample size formulae for multilevel logistic regression, which can handle only models with one predictor. Therefore, and because the four predictors were uncorrelated with each other due to the randomization, we ran models with IN or OUT as only predictor to obtain reasonable guesses of the true parameter values β_0, β_4, β_5, and σ^2_{u0}, which are needed to derive the optimal design. See Table 3 for the parameter estimates obtained with second order Penalized Quasi Likelihood. We performed a bias-corrected parametric bootstrap for the two models with one predictor variable and found that there was hardly any bias in the parameter estimates given in Table 3, but this will not necessarily be the case for all data sets.

TABLE 3
Predictors of Smoking Behavior

Variable	*Estimate*	*Standard Error*	*Z*	*p Value*
All predictor variables				
Constant	$\hat{\beta}_0 = -9.632$	1.559	–6.178	.000
Pretreatment smoking behavior	$\hat{\beta}_1 = 2.920$	0.166	17.597	.000
Age	$\hat{\beta}_2 = 0.555$	0.132	4.208	.000
Condition				
Program (IN)	$\hat{\beta}_4 = -0.084$	0.205	–0.410	.682
Tailor (OUT)	$\hat{\beta}_5 = -0.483$	0.205	–2.356	.019
Random school effect	$\hat{\sigma}^2_{u0} = 0.582$	0.165		
IN is only predictor				
Constant	$\hat{\beta}_0 = -2.637$	0.103	25.602	.000
Condition				
Program (IN)	$\hat{\beta}_4 = -0.044$	0.206	0.216	.831
Random school effect	$\hat{\sigma}^2_{u0} = 0.689$	0.168		
OUT is only predictor				
Constant	$\hat{\beta}_0 = -2.637$	0.102	–25.853	.000
Condition				
Tailor (OUT)	$\hat{\beta}_5 = -0.495$	0.204	–2.426	.015
Random school effect	$\hat{\sigma}^2_{u0} = 0.662$	0.165		

Note. The dummy variable IN is coded +0.5 for a pupil receiving the in-school intervention and –0.5 otherwise. The dummy variable OUT is coded likewise.

OPTIMAL SAMPLE SIZES

Optimal Sample Sizes Given a Fixed Total Budget for Sampling and Measuring

Crucial to the planning of a design is the selection of a statistical model for the data analysis. For the smoking prevention intervention in this study, the model given by Equation 3, without the covariate GENDER, will be used, and we describe in this section the optimal design for estimating the treatment effects with this model. Furthermore, an optimality criterion is needed to construct the optimal design. Several statistical optimality criteria are available, and different optimality criteria may lead to different designs. In the smoking prevention intervention, the treatment effects β_4, β_5, and β_6 are of primary interest and should be estimated with minimal standard error because this will lead to the smallest confidence intervals for these regression coefficients and to the highest power of the test for no treatment effect. These standard errors may be calculated as a function of the number of schools (n_2, assuming $n_2/4$ schools per treatment condition), the number of pupils per school (n_1, assumed to be constant across schools), and the variances σ^2_{u0} and σ^2_e, or the intraschool correlation coefficient ρ:

$$SE(\hat{\beta}_4) = SE(\hat{\beta}_5) = SE(\hat{\beta}_6) = 2\sqrt{\frac{n_1\sigma^2_{u0} + \sigma^2_e}{n_1 n_2}} = 2\sqrt{\frac{(\sigma^2_{u0} + \sigma^2_e)[(n_1 - 1)\rho + 1]}{n_1 n_2}}. \quad (5)$$

Again we assume (–0.5, +0.5) coding of the treatment indicators IN, OUT, and INTERAC; zero correlation of these indicators with the pupil-level covariates due to the randomization; and given randomization of schools rather than pupils (see Moerbeek et al., 2001b). This is an important formula because it shows how the standard error is inflated when we use a two-stage sample instead of a simple random sample. Given total error variance $\sigma^2_{u0} + \sigma^2_e$ and total number of pupils $n_1 n_2$, the design effect is defined as the variance (i.e., the squared standard error) obtained with a two-stage sample divided by that obtained with a simple random sample and is equal to

$$\text{design effect} = (n_1 - 1)\rho + 1. \quad (6)$$

Its square root appears in the numerator of Equation 5. If, for instance, $n_1 = 20$ and $\rho = 0.1$, the design effect = $(20 - 1) \times 0.1 + 1 = 2.9$, and thus the standard error obtained with a two-stage sample is $\sqrt{2.9} \approx 1.7$ times as large as that obtained with a

simple random sample. The design effect shows that we cannot treat the observations in a two-stage sample as independent, especially not when $(n_1 - 1)\rho$ is large, irrespective of the value of n_2. Given n_2, the design effect may be used to calculate the effective sample size (i.e., the equivalent total sample size in a simple random sample):

$$n_{effective} = \frac{n_1 n_2}{\text{design effect}}. \tag{7}$$

If, for instance, we have $\rho = 0.1$, $n_1 = 20$, and $n_2 = 10$, we have $n_{effective} = (20 \times 10)/2.9 \approx 69$. So for these values of n_1, n_2, and ρ, a simple random sample with 69 individuals is equivalent to a two-stage with a total of 200 individuals, a sample almost three times as large! Thus, ignoring the intraschool correlation leads to too small sample sizes in the design stage resulting in lack of power and an increase of Type II errors in case of a correct data analysis. Of course, if the data analysis also ignores the intraschool correlation, this will lead to underestimation of standard errors and an increase of Type I errors rather than Type II errors.

Given the total sample size $n_1 n_2$, the standard errors in Equation 5 are minimized by taking the number of schools n_2 as large as possible. The number of schools, however, is limited because enrolling more schools in the intervention study will result in higher costs, and these costs should not exceed the budget that is available for the study. Furthermore, the budget is not only spent on enrolling schools but also on the participation of pupils in the intervention study. It is therefore necessary to derive the optimal number of schools (n_2) and the optimal number of pupils per school (n_1) such that the costs for sampling and measuring will not exceed the maximum budget C that is available for sampling and measuring. The costs may be calculated by multiplying the total number of pupils $n_1 n_2$ by the costs c_1 per pupil and adding the total costs at the school level, which are calculated by multiplying the number of schools n_2 by the costs c_2 per school. We then have the following precondition for minimizing the standard errors in Equation 5:

$$c_1 n_1 n_2 + c_2 n_2 \leq C. \tag{8}$$

The optimal sample sizes can then be shown to be equal to

$$\begin{aligned} n_1 &= \sqrt{\frac{\sigma_e^2 c_2}{\sigma_{u0}^2 c_1}} = \sqrt{\frac{(1-\rho)c_2}{\rho c_1}}, \text{ and} \\ n_2 &= \frac{C}{c_2 + \sqrt{\frac{\sigma_e^2}{\sigma_{u0}^2} c_1 c_2}} = \frac{C}{c_2 + \sqrt{\frac{1-\rho}{\rho} c_1 c_2}}, \end{aligned} \tag{9}$$

and the standard errors given the optimal sample sizes in Equation 9 are equal to

$$SE(\hat{\beta}_4) = SE(\hat{\beta}_5) = SE(\hat{\beta}_6) = 2\frac{\sqrt{\sigma^2_{u0}c_2} + \sqrt{\sigma^2_e c_1}}{\sqrt{C}}; \quad (10)$$

see Cochran (1977) and Moerbeek et al. (2000). It follows from Equation 9 that the optimal n_1 and n_2 depend on the costs c_1 and c_2 and the intraschool correlation coefficient ρ. The budget C does not have an effect on the optimal n_1, but the optimal number of schools increases if the budget increases. With increasing the budget C, it is always more efficient to increase n_2 than n_1, and therefore, n_1 does not vary with C, given cluster randomization of course. See Equation 5: The effect of σ^2_{u0} on the standard error is reduced only by increasing n_2, not by increasing n_1, whereas the effect of σ^2_e is reduced by increasing $n_1 n_2$. So in total, increasing n_2 is more efficient than increasing n_1 if C increases.

In general the optimal sample sizes calculated from Equation 9 are real values and have to be rounded off to integer values such that the budget is not exceeded, and the number of schools in this study are assumed to be a multiple of four because there are four treatment groups that are assumed to consist of equal numbers of schools. Rounding off to integer values may lead to a somewhat higher standard error than calculated from Equation 10.

In deriving Equations 5, 9, and 10, we assumed that the four treatment groups had equal numbers of clusters (i.e., the number of clusters is divisible by four) and that the cluster sizes did not vary. Moreover we assumed that treatment condition was uncorrelated with the other predictor variables, which may be achieved by combining randomization with large sample sizes or prestratification on the covariates. Such a balanced design was chosen because (a) it minimizes the standard error of the treatment effect estimator and (b) it leads to relatively simple formulae for the optimal sample sizes and the standard error of the treatment effect estimator. In reality, cluster sizes most often vary (e.g., if the clusters are families or existing classes within schools). In that case, equal cluster sizes may be achieved by taking a subsample from each cluster equal to the optimal number of individuals per cluster as given by Formula 9, provided that the actual cluster sizes are larger than the optimal cluster size. If this raises ethical or practical concerns, we may choose to include all individuals in the sampled clusters. In that case, we can substitute the mean cluster size in Formulae 5, which gives some underestimation of the standard error of the treatment effect estimator, depending on the amount of variability in cluster sizes (see Manatunga, Hudgens, & Chen, 2001). Also, the assumption of the number of clusters divisible by four may be relaxed by using an n_2 divisible by four to obtain an upper limit for the standard error of the treatment effect estimator from Equation 5 and actually recruiting an many clusters as the budget C allows because that results in a lower standard error.

In some studies, a limited number of possible schools may be available for the intervention study. Then, it can be shown that the optimal number of schools as given by Equation 9 should be used if this limited number is larger than the optimal number of schools, and the number of pupils per school and standard error can be calculated from Equations 9 and 10, respectively. The limited number of schools should be used if this number is smaller than the optimal number of schools given by Equation 9, and then the optimal number of pupils per school and the standard error can be calculated from Equations 8 and 5, respectively. Thus, the number of pupils per school will in this case be determined by the budget only (i.e., recruit as many pupils as the budget allows). Adding pupils, however, will hardly decrease the standard error of the treatment effect estimator if the number of pupils per school is already high. A similar method can be applied when the number of pupils per school is bounded to a maximum.

Moerbeek et al. (2001a) gave optimal designs for multilevel models with binary responses, one treatment factor, and no covariates. For estimation with second order Penalized Quasi Likelihood, the optimal samples sizes are equal to those for the linear case as given in Equation 9 with σ_e^2 replaced by the following term: $\delta^2 = \frac{1}{2}(4 + e^{\beta_0+\frac{1}{2}\beta_4} + e^{\beta_0-\frac{1}{2}\beta_4} + e^{-\beta_0+\frac{1}{2}\beta_4} + e^{-\beta_0-\frac{1}{2}\beta_4})$, which can be shown to be equal to at least 4. Here β_0 is the intercept and β_4 is the regression coefficient associated with the variable IN, which has coding scheme [–0.5, +0.5]. Their values as well as the value of σ_{u0}^2, have to be known or inferred from previous studies or notions about the smallest relevant treatment effect to calculate the optimal design. Remember that for the linear multilevel model only the values of the variance components were needed to calculate the optimal design. For the variable OUT, β_4 has to be replaced by β_5. The study by Moerbeek et al. (2001a) showed that for second order Penalized Quasi Likelihood, the standard errors as calculated from Equations 5 and 10 hold approximately if they are multiplied by a factor $\sqrt{1.2}$ and if σ_e^2 is replaced by δ^2. The formula for δ^2 is more complicated for models with two treatment factors and covariates. Therefore, the optimal design will be approximated by using parameter estimates from the model in which the variable of interest (IN or OUT) is the only predictor variable. This approximation has been shown to work well when this variable is uncorrelated with the other predictor variables, which is achieved by randomization, and when the variance σ_{u0}^2 is equal to zero (Hsieh, Bloch, & Larsen, 1998), and future research has to show to what extent this approximation holds for multilevel models.

Optimal Sample Sizes Given a Minimally Required Statistical Power

Until now, the budget was given and the optimal sample sizes and standard error were calculated. It is, however, also possible to start with a maximum allowable

standard error because the standard error determines the power of the test of the treatment effect and the width of the confidence interval for the treatment effect. The necessary budget C can then be calculated from Equation 10 and the optimal sample sizes n_1 and n_2 follow from Equation 9. If the available budget is much smaller than the necessary budget, then maybe the study should not be run because power would be too low.

The maximum allowable standard error may be calculated such that the test on treatment effect has a given power level. The following formula relates standard error to the power $1-\gamma$, significance level α, and the minimal relevant deviation Δ of the treatment effect β_4, β_5, or β_6 from zero:

$$SE(\hat{\beta}_4) = SE(\hat{\beta}_5) = SE(\hat{\beta}_6) = \frac{\Delta}{z_{1-\alpha/2} + z_{1-\gamma}}. \quad (11)$$

$z_{1-\alpha/2}$ and $z_{1-\gamma}$ are the $100(1-\alpha/2)\%$ and $100(1-\gamma)\%$ standard normal deviates. Formula 11 holds for two-sided tests; for a one sided test, $z_{1-\alpha/2}$, of course, has to be replaced by $z_{1-\alpha}$. In calculating the power, it is assumed that the Wald test statistic for the treatment effect has a standard normal distribution under the null hypothesis, and this assumption seems acceptable for continuous outcomes and also for binary outcomes if second order Penalized Quasi Likelihood is used for the data analysis (Moerbeek et al., 2002).

The value of Δ may be difficult to specify. For continuous outcomes, we may use the effect size (ES) for two independent means as given by Cohen (1992):

$$\text{effect size} = \frac{\text{mean(E)} - \text{mean(C)}}{\text{standard deviation}}, \quad (12)$$

where mean(E) and mean(C) are the mean outcomes in the experimental and control group, respectively. The ES is the degree to which the null hypothesis of no difference between these two means is believed to be false. It is scale free and continuous, and ranges upward from zero, and ES = 0 corresponds with H_0. Cohen used 0.2, 0.5, and 0.8 for small, medium, and large ES, respectively. A medium ES represents an effect likely to be visible to the naked eye of a careful researcher. The numerator of Equation 12 is simply equal to the treatment effect (either β_4, β_5, or β_6) because of the [–0.5, +0.5] coding scheme used; see the Description of the Intervention Study Used As Example section. The standard deviation is equal to $\sqrt{\sigma^2_{u0} + \sigma^2_e}$ because the residual variance of the outcomes is $\sigma^2_{u0} + \sigma^2_e$ for both the experimental and control group, see the Description of the Intervention Study Used As Example section. Once an ES is chosen and the values of the variance components are known

or estimated from previous studies, Δ follows from Equation 12, $\Delta = \mathrm{ES} \times \sqrt{\sigma^2_{u0} + \sigma^2_e}$, and may be substituted into Equation 11 to calculate the maximum allowable standard error.

For binary outcomes, we may use the odds ratio to get a value of Δ. The odds ratio for the treatment effect β_4 is equal to $\exp(\beta_4)$ and likewise for β_5 and β_6. Once a minimal relevant odds ratio is specified, we can calculate $\Delta = \ln(\text{odds ratio})$ and substitute into Equation 11 to calculate the maximum allowable standard error. Translating this standard error into the budget C and sample sizes n_1 and n_2 requires specification of β_0 and σ^2_{u0}, of course.

OPTIMAL SAMPLE SIZES FOR THE DUTCH SMOKING PREVENTION INTERVENTION

Specification of Costs and Budget

Prior to the calculation of the optimal sample sizes, we need to specify the costs at the pupil and school level and the total budget C. These will all be based on the actual study by Ausems et al. (2002), that is, the initial sample consisted of 160 schools and on average 25 pupils per school. Due to 12.5% dropout at the school level and 4% at the pupil level, their actual study size was 140 schools and on average 24 pupils per school. Although data on the dropped-out schools and pupils are not available at both pretest and posttest, they do demand part of the budget and therefore this dropout has to be taken into account when calculating optimal designs.

The following costs were calculated for the smoking prevention intervention. The costs are given in European Euros, which are denoted €. The average costs for including a pupil are € 4.55. The average costs for including a school are € 119.10. Note that we use the costs averaged over the treatment conditions because formulae on optimal sample sizes that take into account costs dependent on treatment condition are not yet available for nested data (see Schouten, 1999, for nonnested data). The total costs of the intervention study were equal to C = € 36363.63, which is a reasonable quantity for Dutch intervention studies and will be used as budget C for calculating the optimal design. Figure 1 shows the maximum number of pupils per school that can be enrolled in the study as a function of the number of schools. This relationship follows from Equation 8 by substitution of c_1 = € 4.55, c_2 = € 119.10, and C = € 36363.63, which gives € 4.55 × $n_1 n_2$ + € 119.10 × n_2 = € 36363.63 and subsequently solving for n_1, which gives n_1 = (€ 36363.63 – € 119.10 × n_2)/(€ 4.55 × n_2). As is obvious, the number of pupils per school decreases if the number of schools increases.

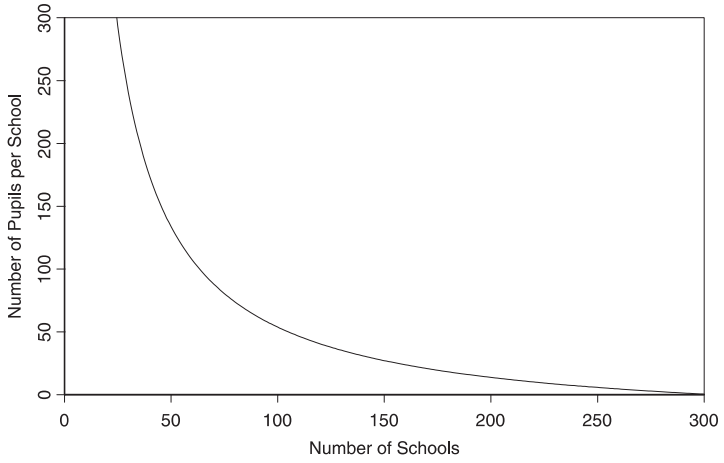

FIGURE 1 Number of pupils per school as a function of the number of schools, given *C* and Formula 8.

Optimal Experimental Designs for the Outcome Variable Attitude

The optimal design for the outcome variable ATTITUDE was calculated for $c_1 = €$ 4.55, $c_2 = €$ 119.10, $C = €$ 36363.63, and $\sigma^2_{u0} = 3.349, \sigma^2_e = 44.952$ (see Table 2 for these values of the variance components). The optimal designs for the in-school intervention, the out-of-school intervention, and the interaction term are equal because there are equal numbers of schools in each treatment group and the true (population) values of the regression coefficients β_4, β_5, and β_6 are not necessary for calculating the optimal design; see Equations 5 and 10. In calculating the optimal design, we used the fact that the variables are uncorrelated with each other and the covariates due to the (–0.5, +0.5) coding scheme that was used and the randomization procedure.

Figure 2 shows the standard error of the treatment effects as a function of the number of pupils per school (upper half) and the number of schools (lower half), if the complete budget *C* is used. To draw the upper half of this figure, we first obtain the number of schools for each value of the number of pupils per school from Figure 1. This gives us pairs (n_1, n_2) to plug in Equation 5 to get the standard error of the treatment effect, which we can then draw as a function of n_1. The lower half of Figure 2 is drawn likewise. The shape of the upper half of Figure 2 can be understood by writing the standard error as $SE = (\sigma^2_{u0} + \sigma^2_e / n_1) / n_2$. Increasing n_1 results in a smaller numerator but also a smaller denominator because n_2 decreases if n_1 increases. The effect of the numerator dominates for small n_1, but the effect of

the denominator dominates for large n_1. The effects outweigh each other for the number of schools between 150 and 200; see the flat part in the lower half of Figure 2. According to Cochran (1977, p. 281) such a flat part occurs in most practical situations.

The optimal number of pupils per school can be calculated from Equation 9: $n_1 = \sqrt{(\sigma_e^2 c_2)/(\sigma_{u0}^2 c_1)} = \sqrt{(44.952 \times 119.10)/(3.349 \times 4.55)} = 18.7$. The optimal n_2 follows from Figure 1 and is equal to $n_2 = 178.0$. Rounding off to integer values such that the budget is not exceeded and taking n_2 a multiple of four (because there

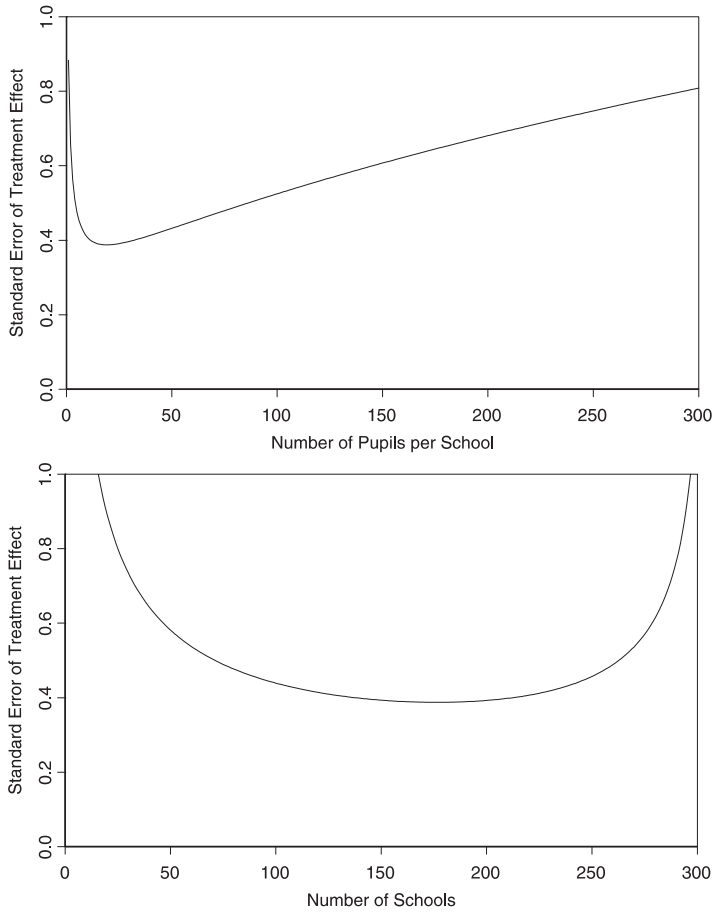

FIGURE 2 Standard error of the treatment effect as a function of the number of pupils per schools (top) and the number of schools (bottom) for the continuous outcome variable ATTITUDE, given the restriction in Figure 1.

are four treatment groups that have to consist of equal numbers of schools) results in $n_1 = 19$, $n_2 = 176$. The total costs are then $c_1 n_1 n_2 + c_2 n_2 = €\ 4.55 \times 19 \times 176 + €\ 119.10 \times 175 = €\ 36176.8$. Due to the dropout, the n_1 and n_2 decrease by 4% and 12.5%, respectively, which gives $n_1 = 19 \times (1 - 0.04) = 18.2$, and $n_2 = 176 \times (1 - 0.125) = 154$. To get the standard error, we substitute these sample sizes and the values of the variance components into Equation 5:

$$SE(\hat{\beta}_4) = 2\sqrt{(n_1\sigma^2_{u0} + \sigma^2_e)/(n_1 n_2)}$$
$$= 2\sqrt{(18.2 \cdot\ 3.349 \cdot\ 44.952)/(18.2 \cdot\ 154)} = 0.389.$$

See Design 1 in Table 4.

Design 2 is the design that has $n_2 = 160$ schools, which is the number of schools used in the cost calculation. The standard error for Design 2 is almost equal to that for the optimal design because the number of schools of both designs are between 150 and 200, which is the flat part in the lower plot of Figure 2.

The number of schools for Design 1 and 2 is very large. Suppose that only $n_2 = 40$ schools are willing to participate in the intervention study. Then $n_1 = 173.6$ pupils per school can be enrolled if we want to use the whole budget C, as follows

TABLE 4
Sample Sizes and Standard Error of the Treatment Effect for Several Designs

Design	*Requirements*	n_1	n_2	*Costs* €	*Standard Error*
Linear Multilevel Model With Outcome Variable ATTITUDE					
1	Optimal design	19	176	36176.80	0.389
2	$n_2 = 160$	23	160	35800.00	0.392
3a	$n_2 = 40$	173	40	36250.00	0.643
3b	$n_2 = 40$, $n_1 = 25$	25	40	9314.00	0.773
4a	$n_2 = 20$	373	20	36325.00	0.891
4b	$n_2 = 20$, $n_1 = 25$	25	20	4657.00	1.093
5	$SE = 0.43$	19	144	29599.20	0.430
Multilevel Logistic Model With Outcome Variable SMOKE and Predictor Variable IN					
1	Optimal design	25	156	36324.60	0.218
2	$n_2 = 160$	23	160	35800.00	0.220
3a	$n_2 = 40$	173	40	36250.00	0.323
3b	$n_2 = 40$, $n_1 = 25$	25	40	9314.00	0.430
4a	$n_2 = 20$	373	20	36325.00	0.441
4b	$n_2 = 20$, $n_1 = 25$	25	20	4657.00	0.608
5	$SE = 0.28$	26	92	21840.80	0.281

Note. The standard errors are calculated for a drop out rate of 4% at the pupil level and 12.5% at the school level. The variable IN is coded +0.5 for a pupil receiving the in-school intervention and –0.5 otherwise. The variable OUT is coded likewise, and the variable INTERAC = 2*IN*OUT, therefore, it has values –0.5 and +0.5 as well.

from Figure 1. We round off to $n_1 = 173$ and the total costs are then $c_1 n_1 n_2 + c_2 n_2 =$ € 4.55 × 173 × 40 + € 119.10 × 40 = € 36250. The dropout rates of 4% and 12.5% at the pupil and school level, respectively, give us $n_1 = 173 \times (1 - 0.04) = 166.1$, and $n_2 = 40 \times (1 - 0.125) = 35$. Substitution into Equation 5 yields $SE(\hat{\beta}_4) = 2\sqrt{(n_1\sigma^2_{u0} + \sigma^2_e)/(n_1 n_2)} = 2\sqrt{(166.1 \times 3.349 + 44.952)/(166.1 \times 35)} = 0.643$ (see Design 3a in Table 4), which is substantially higher than for Designs 1 and 2. This number of pupils per school, however, is unrealistically high because the intervention aims at 11- and 12-year-old pupils and such a large number of pupils is seldom available in Dutch elementary schools. Design 3b is a design for which $n_1 = 25$ pupils per school and $n_2 = 40$ schools are available. The standard error is then equal to 0.773, which is twice as large as for the optimal design. The corresponding costs are of course smaller.

Design 4a and 4b are the designs for which the number of schools is even lower ($n_2 = 20$), and the available number of pupils per school is limited by the budget, or limited to 25, respectively. It can be seen that the standard errors of these designs are substantially higher than those for Designs 1, 2, and 3. The standard errors and used budget for Designs 3b, 4a, and 4b can be calculated similar to those of Designs 3a, and the reader is encouraged to verify the values as given in Table 4.

Now suppose that we want to detect a small (0.2) ES of the in-school intervention with power $1 - \gamma = 0.9$ at a significance level of $\alpha = 0.05$ in a two-sided test. We have $z_{1-\alpha/2} = z_{0.975} = 1.96$, $z_{1-\gamma} = z_{0.9} = 1.28$, and $\sigma^2_{u0} + \sigma^2_e = 48.301$. Then $\Delta = \text{ES} \times \sqrt{\sigma^2_{u0} + \sigma^2_e} = 0.2 \times \sqrt{48.301} = 1.39$. Substitution into Equation 11 gives $SE(\hat{\beta}_4) = (\Delta)/(z_{1-\alpha/2} + z_{1-\gamma}) = 1.39/(1.96 + 1.28) = 0.43$. The costs for the schools and pupils that do not drop out follows from substitution into Equation 10 $SE(\hat{\beta}_4) = 2(\sqrt{\sigma^2_{u0} c_2} + \sqrt{\sigma^2_e c_1})/(\sqrt{C})$, which gives $0.43 = 2(\sqrt{3.349 \times 119.10} + \sqrt{44.952 \times 4.55})/(\sqrt{C})$ and a solution $C = 25411.45$. The optimal n_1 follows from Equation 9 and is equal to that of Design 1 because it does not depend on the budget *C:* $n_1 = 18.7$. The optimal n_2 follows from Figure 1: $n_2 = 124.5$. These are the optimal sample sizes if we had no dropout. Due to the dropout, these sample sizes have to be increased: $n_1 = 18.7/(1 - 0.04) = 19.5$, and $n_2 = 124.5/(1 - 0.125) = 142.3$. Rounding off to integers such that n_2 is a multiple of four, gives $n_1 = 19$ and $n_2 = 144$. The total costs are then equal to $c_1 n_1 n_2 + c_2 n_2 =$ € 4.55 × 19 × 144 + € 119.10 × 144 = € 29599.20. The standard error is equal to

$$SE(\hat{\beta}_4) = 2\sqrt{(0.96 \cdot n_1\sigma^2_{u0} + \sigma^2_e)/(0.96 \cdot n_1 \cdot 0.875 \cdot n_2)}$$
$$= 2\sqrt{(0.96 \cdot 19 \cdot 3.349 + 44.952)/(0.96 \cdot 19 \cdot 0.875 \cdot 144)} = 0.430.$$

This is Design 5.

In Figure 3, the power for two-sided tests with significance level $\alpha = 0.05$ is plotted as a function of the treatment effect. Figure 3 shows that almost equal power is obtained by Design 1 and 2 and a somewhat lower power by the less expensive Design 5. The powers for Designs 3 and 4, however, are much lower, and these designs are only able to detect large treatment effects with high power.

Optimal Experimental Designs for the Outcome Variable Smoking Behavior

Optimal sample sizes for the multilevel logistic model with the variable OUT as only predictor were calculated for the values $\hat{\beta}_0 = -2.637$, $\hat{\beta}_4 = -0.495$, $\hat{\sigma}^2 = 16.475$, and $\hat{\sigma}^2_{u0} = 0.662$ (see Table 3 for these parameter values). The standard error of the treatment effect is plotted as a function of the number of pupils per school and the number of schools in Figure 4. Note that the shapes of the two curves are similar to those in Figure 2.

The optimal number of pupils per school is calculated from Equation 9 with σ^2_e replaced by δ^2: $n_1 = \sqrt{(\delta^2 c_2)/(\sigma^2_{u0} c_1)} = \sqrt{(16.475 \times 119.10)/(0.662 \times 4.55)} = 25.5$.

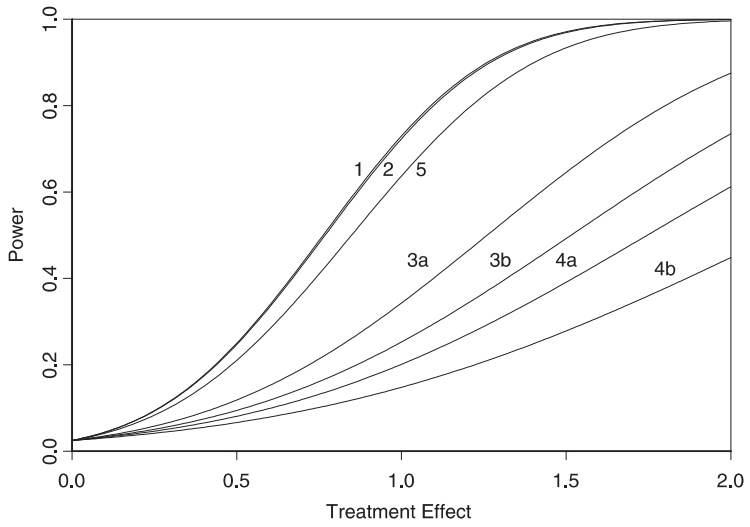

FIGURE 3 Power of two-sided tests as a function of the treatment effect for the designs of Table 4 for the continuous outcome variable ATTITUDE. The designs are indicated by their numbers.

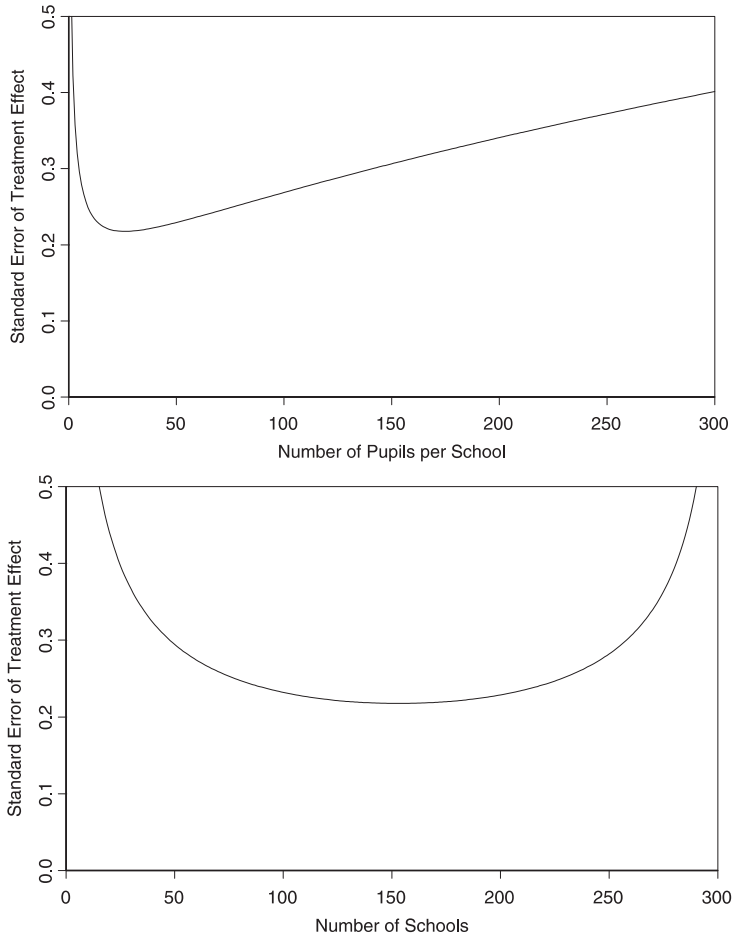

FIGURE 4 Standard error of the treatment effect as a function of the number of pupils per schools (top) and the number of schools (bottom) for the binary outcome variable SMOKE given the restriction in Figure 1.

The optimal n_2 follows from Figure 1: $n_2 = 154.7$. Rounding of to integer values such that n_2 is a multiple of four gives $n_1 = 25$, $n_2 = 156$ and costs $c_1 n_1 n_2 + c_2 n_2 =$ € 4.55 × 25 × 156 + € 119.10 × 156 = € 36324.60. We have a dropout of 4% and 12.5% at the pupil and school level, respectively, resulting in $n_1 = 25 \times (1 - 0.04) = 24$, and $n_2 = 156 \times (1 - 0.125) = 136.5$. The standard error is calculated from Equation 5 multiplied by a factor $\sqrt{1.2}$ and with σ_e^2 replaced by δ^2 (see "Optimal Sample Sizes" section):

$$SE(\hat{\beta}_5) = 2\sqrt{(n_1\sigma_{u0}^2 + \delta^2)/(n_1 n_2)}\sqrt{1.2} =$$
$$SE(\hat{\beta}_5) = 2\sqrt{(24 \cdot 0.662 + 16.475)/(24 \cdot 136.5)}\sqrt{1.2} = 0.218.$$

This is Design 1 in the lower half of Table 4. Note that this optimal design differs from that for the continuous outcome variable ATTITUDE, but for both outcome variables the plot of the standard error as a function of the number of schools is very flat when the number of schools is near 160, and designs that yield almost minimal standard error for both outcome variables are therefore available.

The sample sizes and costs for Designs 2, 3a, 3b, 4a, and 4b are equal to those for the outcome variable ATTITUDE, and the user is encouraged to verify the values in Table 4. Note that the standard error for Design 2 is almost equal to that obtained with the optimal Design 1 because the number of schools of both designs are between 125 and 175, which is the flat part in the lower half of Figure 4. The standard error for effect of the out-of-school smoking prevention intervention as given in Table 3 is somewhat smaller than for the optimal design because this optimal design was calculated with the use of an approximate method, which may somewhat underestimate or overestimate the standard error; for details, see Moerbeek et al. (2001a).

Suppose that we want to detect an odds ratio $\exp(\beta_5)$ equal to 2.5 with power $1 - \gamma = 0.9$ at a significance level of $\alpha = 0.05$ in a two-sided test. Then $\Delta = \ln(2.5) = 0.92$ is the minimal relevant deviation of β_5 from zero. Then from Equation 11, we have $SE(\hat{\beta}_5) = (\Delta)/(z_{1-\alpha/2} + z_{1-\gamma}) = 0.92/(1.96 + 1.28) = 0.28$. The C may be calculated from $SE(\hat{\beta}_5) = 2(\sqrt{\sigma_{u0}^2 c_2} + \sqrt{\delta^2 c_1})\sqrt{1.2}/(\sqrt{C})$, which gives $0.28 = 2(\sqrt{0.662 \times 119.10} + \sqrt{16.475 \times 4.55})\sqrt{1.2}/(\sqrt{C})$, with solution $C = 18830.32$. The optimal n_1 is equal to that of Design 1 because it does not depend on the budget C: $n_1 = 24.7$. The optimal n_2 follows from Figure 4 and is equal to 81.3. These are the sample sizes after dropout. To account for the 4% and 12.5% dropout rates at the pupil and school level, respectively, the sample sizes should be increased: $n_1 = 24.7/(1 - 0.04) = 25.7$, and $n_2 = 81.3/(1 - 0.125) = 92.9$. We round off to $n_1 = 26$ and $n_2 = 92$ to have the number of schools a multiple of four. The total costs are then $c_1 n_1 n_2 + c_2 n_2 = €\,4.55 \times 26 \times 92 + €\,119.10 \times 92 = €\,21840.8$, and the standard error is

$$SE(\hat{\beta}_5) = 2\sqrt{(0.96 \cdot n_1\sigma_{u0}^2 + \delta^2)/(0.96 \cdot n_1 \cdot 0.875 \cdot n_2)}\sqrt{1.2}$$
$$= 2\sqrt{(0.96 \cdot 26 \cdot 0.662 + 16.475)/(0.96 \cdot 26 \cdot 0.875 \cdot 92)}\sqrt{1.2} = 0.281.$$

This is Design 5 in Table 4.

Figure 5 gives the power of two-sided tests for the designs given in Table 4 as a function of the treatment effect. The power of the optimal design is almost equal to

that of Design 2. A somewhat smaller power is obtained with the less expensive Design 5. The power for Designs 3 and 4 is much lower, and these designs should not be used if the study aims at detecting a small treatment effect because power would be too low.

Finally, it should be noted that for the variable IN, which is associated with the in-school intervention, the derivation of the optimal design is almost equal to that for the variable OUT and is not given here.

DISCUSSION AND CONCLUSIONS

In this article, we have demonstrated how optimal sample sizes for experiments in which individuals are nested within clusters can be calculated. As an example, a school-based smoking prevention intervention was used. Two treatment factors with two levels each were evaluated in the experiment, resulting in four treatment groups. The formulae presented in the "Optimal Sample Sizes" section may, however, also be used if there is only one treatment factor with two levels.

The optimal designs presented in this article are for randomization at the school level. As follows from Moerbeek et al. (2000, 2001a, 2001b), a more efficient design can be achieved by randomization at the pupil level, especially when the intraschool correlation and the number of pupils per school are large. In the school-based smoking prevention intervention, which was used as an example in

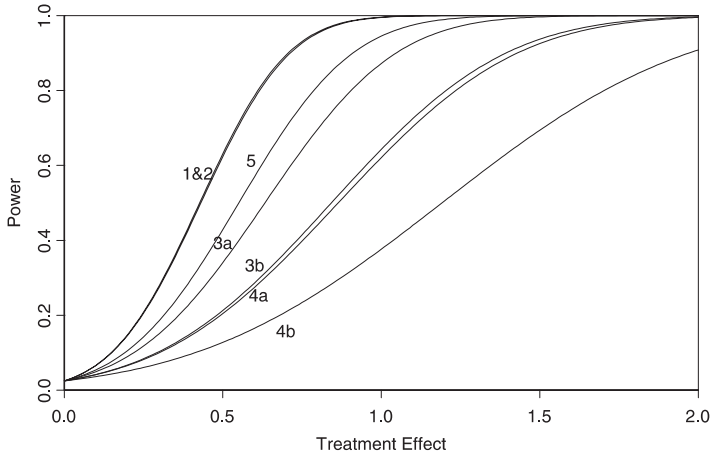

FIGURE 5 Power of two-sided tests as a function of the treatment effect for the designs of Table 4 for the binary outcome variable SMOKE. The designs are indicated by their numbers. Note that the curves for Designs 1 and 2 almost overlap.

this article, this was not possible because the in-school intervention consisted of lessons on different aspects of smoking and had to be delivered class wise. The out-of-school intervention, which consisted of personalized tailored letters, was delivered at the school level as well to avoid treatment group contamination. In other studies, however, randomization at the person level is feasible, for example, in double-blind multicenter clinical trials for the evaluation of, for instance, antibiotics or antihypertensiva. Formulas for calculating the optimal sample sizes at both levels, given person randomization, are very similar to those presented in this article (except that the budget C then may affect n_1 as well as n_2) and can be found in Moerbeek et al. (2000, 2001a, 2001b).

Our future research may focus on the case in which the costs c_1 and c_2 as well as the variance components depend on the treatment condition. Also, it should be noted that in this article, the optimal design was calculated ignoring dropout and the optimal sample sizes thus obtained were afterward adjusted for dropout. Our future research will take dropout into account while calculating the optimal design rather than adjusting the sample sizes afterward.

ACKNOWLEDGMENTS

The Dutch smoking prevention intervention is part of a European project, the Octopus project, which is supported by financial grants from the European Commission (Soc 96 200568 05FO2/Soc 97 20249105F02).

REFERENCES

Ausems, M., Mesters, I., Van Breukelen, G., & De Vries, H. (2002). Short-term effects of a randomized computer-based out-of-school smoking prevention trial aimed at elementary schoolchildren. *Preventive Medicine, 34,* 581–589.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T., Jr. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L Programs.* Chicago: Scientific Software International.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

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

Donner, A., Birkett, N., & Buck, C. (1981). Randomization by cluster. Sample size requirements and analysis. *American Journal of Epidemiology, 114,* 906–914.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., et al. (2000). *A user's guide to MLwiN, Version 2.1c.* London: Institute of Education.

Hedeker D., & Gibbons, R. D. (1996a). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine, 49,* 157–176.

Hedeker D., & Gibbons, R. D. (1996b). MIXREG: A computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine, 49,* 229–252.

Hox, J. J. (2002). *Multilevel analysis techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hsieh, F. Y. (1988). Sample size formulae for intervention studies with the cluster as the unit of randomization. *Statistics in Medicine, 8,* 1195–1201.

Hsieh, F., Bloch, D., & Larsen, M. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17,* 1623–1634.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modelling.* London: Sage.

Little, R. J. A., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute.

Longford, N. T. (1995). *Random coefficient models.* Oxford, England: Clarendon.

Manatunga, A. M., Hudgens, M. G., & Chen, S. (2001). Sample size estimation in cluster randomised studies with varying cluster size. *Biometrical Journal, 43,* 75–86.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

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

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001a). Optimal experimental designs for multilevel logistical models. *The Statistician, 50,* 17–30.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001b). Optimal experimental designs for multilevel logistical models with covariates. *Communications in Statistics—Theory and Methods, 30,* 2683–2697.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003a). A comparison between traditional and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology, 56,* 341–350.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2003b). A comparison of estimation methods for mulitlevel logistic models. *Computational Statistics, 18,* 19–37.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185.

Raudenbush, S. W., & Bryk, A. S (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rodríguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society. Series A. Statistics in Society, 164,* 339–355.

Schouten, H. J. A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine, 18,* 87–91.

Siddiqui, O., Hedeker, D., Flay, B. R, & Hu, F. B. (1996). Intraclass correlation estimates in a school based smoking prevention study. *American Journal of Epidemiology, 144,* 425–433.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two- level research. *Journal of Educational Statistics, 18,* 237–259.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* London: Sage.