A review of four methods for the analysis of multilevel experimental data

Mirjam Moerbeek[1], Gerard J.P. Van Breukelen[2], Martijn P.F. Berger[2]

Abstract

This paper reviews three traditional methods, namely the fixed effects model, the disaggregated data model, and the aggregated data model, for the analysis of experimental data where individuals are nested within clusters, and compares these methods with the multilevel (mixed effects) model. The comparison is made for continuous outcomes, and is based on the estimator of the treatment effect and its variance, since these usually are of main interest in experiments. When the results of the experiment have to be valid for a larger population of clusters, the clusters in the experiment have to represent a random sample from this population and the multilevel model is preferably used for the analysis of the data. The three traditional methods have the same treatment effect estimator as the multilevel model if cluster sizes do not vary and there are no covariates. The variance of this estimator, however, may be underestimated or overestimated by the fixed effects and disaggregated data models, resulting in an inflated type I or type II error rate for the test on treatment effect, respectively. The aggregated data model may be a good alternative to multilevel analysis if the cluster sizes do not vary and the model does not contain covariates.

Key words: multilevel model, fixed effects model, disaggregated data model, aggregated data model, mixed effects ANOVA, mixed effects regression

Acknowledgments: We wish to thank Brian R. Flay for his permission to use the TVSFP data, which were collected with funding from the National Institute of Drug Abuse, Grant 1-R01-DA03468 to Brian R. Flay, W.B. Hansen, and C.A. Johnson.

[1] National Institute of Public Health and the Environment, Laboratory of Health Effects Research, PO Box 1, 3720 BA Bilthoven. E-mail: Mirjam.Moerbeek@RIVM.NL. Phone: 030-2742214

[2] Maastricht University, Department of Methodology and Statistics, PO Box 616, 6200 MD Maastricht. E-mail: Gerard.vBreukelen@STAT.UNIMAAS.NL and Martijn.Berger@STAT.UNIMAAS.NL

1 Introduction

Experiments are developed to compare different treatments in terms of outcome variables measuring the health or behaviour of individuals. In this paper we will focus on experiments where the data have a nested or hierarchical structure in which individuals are nested within clusters. For example, in the clinical trial analyzed by Hedeker, Gibbons, and Davis (1991) on the effect of different antipsychotics on the mental health, patients were nested within centers. In the trial by Bass, McWinney, and Donner (1986) where a new approach for the detection and managing of hypertension was studied, patients were nested within family practices. In the study by Sommer *et al.* (1986) on the effect of vitamin A supplementation on childhood mortality in northern Sumatra, children were nested within villages, and in the smoking cessation intervention by Hedeker, McMahon, Jason, and Salina (1994) employees were nested within worksites. Outcomes of individuals within the same cluster are likely to be correlated, i.e. there will be intra-cluster correlation.

The effect of an active treatment can be estimated with a regression model, in which the outcome variable is regressed on treatment condition and relevant covariates, or with an analysis of variance (ANOVA) model. In the literature, several types of regression models are being used for multilevel experimental data. Three traditional regression models are the fixed effects model, the disaggregated data model, and the aggregated data model. In the fixed effects model, clusters are treated as fixed and their differences are taken into account by dummy coding in the regression model. In the disaggregated data model individuals are the unit of analysis and their nesting within clusters, i.e. the dependency among the outcomes of individuals within a cluster, is ignored. The aggregated data model is based upon aggregation of data within the same treatment condition to the cluster level, and clusters are thus the unit of analysis.

In the multilevel model (Goldstein, 1995; Hox, 1994; Kreft and De Leeuw, 1998; Snijders and Bosker, 1999) individuals are treated as the unit of analysis, but the dependency of outcomes of individuals nested within the same cluster is also taken into account. This model is also referred to as mixed effects regression, random coefficient model (Longford, 1995) or hierarchical linear model (Bryk and Raudenbush, 1992), and assumes the clusters and individuals to represent random samples from corresponding populations. Under this assumption cluster and person effects must be treated as random effects in the regression model, while treatment condition and covariates may be included as fixed effects.

Ideally, the aim of experiments in a hierarchically structured population should be to produce results which are not only valid for the clusters involved in the experiment, but also for a larger population of clusters. In that case the clusters involved in the experiment have to represent a random sample from a population of clusters, and multilevel analysis is the best method of analysis since it includes the clusters as random effects in the statistical model. There may be practical reasons for treating clusters as fixed, for instance when the number of clusters in the experiment is very small, for details see Senn (1998). In this paper, however, we will focus on the situation where the clusters involved in the experiment may be considered a random sample from a much larger population of clusters.

The multilevel model is more complex than the more traditional models, and consequently investigators may still want to use these traditional models, even if they want to generalize the results from their experiment to all clusters in the population. Therefore a comparison between the traditional models and the multilevel model for experiments in hierarchically structured populations is relevant. In this paper, the relationship between the four models will be given and it will be shown under which circumstances the traditional methods are acceptable, and when and how they may lead to poor results. This will be done both in terms of regression and ANOVA models. ANOVA is familiar to many researchers and works well when cluster sizes do not vary, but encounters more difficulties with varying cluster sizes. The comparison made in this paper is based on analytical expressions for the estimator of the treatment effect and its variance, since these are of main interest in experiments, and is done for models with continuous outcomes, two levels of nesting (individuals within clusters), and with randomization at either level. Clusters will be randomly allocated to the treatment conditions for randomization at the cluster level, and all individuals within each cluster receive the same treatment condition. For randomization at the individual level, half of the individuals within a cluster will be randomized to the treatment group while the others will be randomized to the control group.

Part of the comparison has already been made by others, but has been published fragmentarily in various papers. In the present paper, these results will be presented systematically, and some gaps in knowledge will be filled up. From the literature it is known that:

1. Multilevel analysis is equivalent to a mixed model ANOVA when cluster sizes do not vary (Raudenbush, 1993).
2. The fixed effects model gives a smaller variance of the treatment effect estimator than the multilevel model (Senn, 1998; Gould, 1998; Jones, Teather, and Lewis, 1998) when randomization is done at the individual level and there is interaction between treatment and cluster.
3. For individual level randomization and no interaction between treatment and cluster, the disaggregated data model overestimates the variance of the treatment effect (Parzen, Lipsitz, and Dear, 1998). A similar phenomenon has been shown to occur in longitudinal studies with repeated measurements nested within persons instead of individuals within clusters (Dunlop, 1994).
4. The aggregated data model and the multilevel model lead to the same results if the design is balanced and randomization is done at the cluster level (Hopkins, 1982).
5. The disaggregated data model underestimates the variance of the treatment effect (Hedeker *et al.*, 1994; Longford, 1995) when randomization is done at the cluster level, especially when the number of individuals within clusters and/or the intra-cluster correlation is large (Barcikowski, 1981). Underestimation by the disaggregated data model may also occur in longitudinal studies (Dunlop, 1994).
6. For observational studies with unbalanced designs and covariates, it has been shown that the disaggregated data model underestimates variances of regression coefficients

(Bryk and Raudenbush, 1992), and that the disaggregated and aggregated data model should not be used for the analysis of multilevel data (Aitkin and Longford, 1986).

Again, we want to stress that in this paper multilevel models and more traditional models for *experimental* data with persons within clusters are presented. Multilevel models may also be used for *observational* or *longitudinal* studies. Tutorials on multilevel models for observational studies (Sullivan, Dukes, and Losina, 1999) and longitudinal studies (Burton, Gurrin, and Sly, 1998) have recently appeared.

The remainder of this paper is as follows: in Section 2 an example data set of an experiment in a hierarchically structured population and two different designs for such experiments are given. In Section 3, the multilevel regression model is related to the mixed effects ANOVA model. The fixed effects model, the disaggregated data model and the aggregated data model are presented in Section 4 and related to their corresponding ANOVA models. In Section 5, the four models are used to analyze generated data sets and it is shown that these models lead to different results. This difference in results will be explained using analytical expressions in Section 6. In Sections 3 to 6 we assume balanced designs and no covariates, but in Section 7 these assumptions will be relaxed. In Section 8 some conclusions will be presented. The notation of Goldstein (1995) will be used throughout this paper.

2 Designs and example data set

In principle, randomization and implementation of the two treatments may be done at either level of the hierarchy. So two different designs may be distinguished: Design 1, where randomization is done at the individual level, and Design 2, where randomization is done at the cluster level. The latter is often referred to as cluster randomization. For non-varying cluster sizes we have a sample of n_2 clusters and n_1 individuals per cluster. In Design 1, $\frac{1}{2}n_1$ individuals per cluster are randomized to the control group and the others are randomized to the treatment group; assume n_1 to be even. In Design 2, $\frac{1}{2}n_2$ clusters are allocated to each treatment; assume n_2 to be even, and all individuals within the same cluster receive the same treatment. A graphical representation of these two designs is given in Figure 1 for four clusters. Data on both treatment conditions are available in each cluster for Design 1 and so the interaction between cluster and treatment condition can be estimated. This is not possible in Design 2, where data on only one treatment condition are available per cluster, i.e. the data on the other treatment condition are missing by design. So, individual level randomization is to be preferred to cluster level randomization if treatment by cluster interaction is to be evaluated. Furthermore, randomization at the individual level results in more efficient estimates of the treatment effect (Moerbeek, Van Breukelen, and Berger, 2000). Randomization at this level was done in, for example, the trial analyzed by Hedeker *et al.* (1991). In some experiments, however, randomization at the individual level is not possible and Design 2 will be the only alternative. For example, in the trial by Bass *et al.* (1986) randomization was done at the family practice level since it was recognized that the intervention would not function effectively if some patients in a practice were randomized to

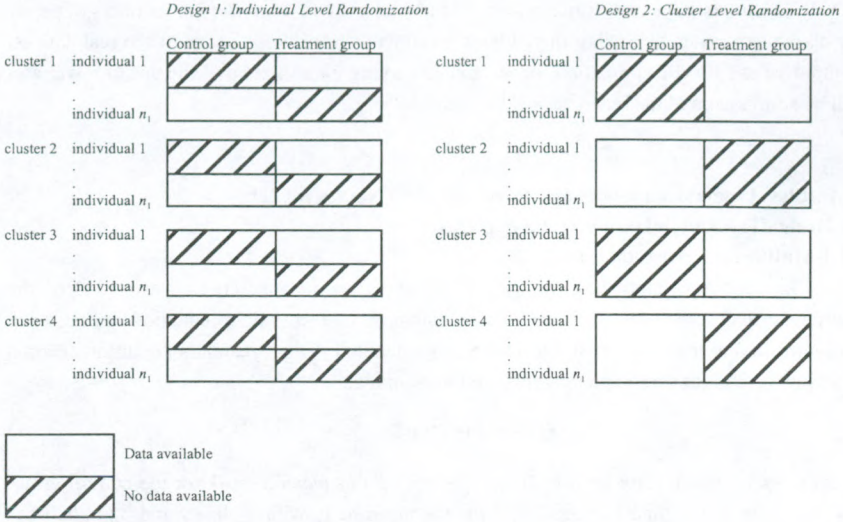

Figure 1. Graphical representation of Design 1 and Design 2.

the treatment group and others not. In the study by Sommer *et al.* (1986), randomization was done at the village level since it was thought to be not acceptable to treat some children in a given village and others not. In the study by Hedeker *et al.* (1994) randomization at the employee level would not have been possible because of treatment group contamination.

The results in this paper will be illustrated using a subset of data from the Television School and Family Smoking Prevention and Cessation Project (TVSFP) (Flay *et al.*, 1988, 1995). This study was designed to test effects of a school-based social-resistance curriculum and a television-based program in terms of tobacco use prevention and cessation. Schools in Los Angeles and San Diego were randomized to one of five treatment conditions: (a) a social-resistance classroom curriculum, (b) a media (television) intervention, (c) a combination of these two, (d) a health-information-based attention-control curriculum, (e) a no-treatment control group. The dependent variable we used in the analyses is the post-intervention Tobacco and Health Knowledge Scale (THKS) score, which was the number of items that a student correctly answered in a seven item questionnaire to assess student tobacco and health knowledge. The data have a nested structure with pupils nested within classes within schools, and varying numbers of pupils per class and classes per school. In this paper we will restrict ourselves to two levels of nesting (pupils within classes) and two treatment conditions (media (television) intervention group and no-treatment control group), and only data from Los Angeles schools are considered. There were seventy classes and 837 pupils who met these conditions.

To illustrate the results in Sections 3 to 6 data sets with non-varying numbers of pupils per class were generated, using the parameter estimates from the analysis of the real data set as input values for the simulation. In Section 7, varying class-sizes and the use of covariates will be addressed and the real data will be analyzed.

3 Multilevel regression model and mixed effects ANOVA model

3.1 Design 1: Randomization at the pupil level

3.1.1 Multilevel regression model

In multilevel modelling, regression equations are formulated for each level of the multilevel data structure, and are then combined into a single equation model. For randomization at the pupil level, the THKS score denoted y_{ij} and treatment condition denoted x_{ij} of pupil i in class j are related by the pupil level model:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}, \tag{1}$$

where e_{ij} is a random error term at the pupil level. In this paper $x_{ij} = -1$ for the control group and $x_{ij} = +1$ for the media group. So, β_{0j} is the mean of y_{ij} within class j and β_{1j} is half the difference in outcome between the two treatments within class j. The intercept and slope may vary across classes, randomly and/or as a function of class level covariates. This section will be restricted to models without any covariate, leaving the inclusion of covariates to Section 7. Thus, $\beta_{0j} = \beta_0 + u_{0j}$, and $\beta_{1j} = \beta_1 + u_{1j}$, where β_0 is the overall mean, β_1 is half the overall treatment effect, and u_{0j} and u_{1j} are random error terms representing the deviation of class j from the overall mean and overall treatment effect, respectively. Substituting β_{0j} and β_{1j} into model (1) yields the single equation model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij}. \tag{2}$$

The random effects u_{0j}, u_{1j}, and e_{ij} are assumed to be independently and normally distributed with zero mean and variances σ^2_{u0}, σ^2_{u1} and σ^2_e, respectively. To compare to ANOVA and for simplicity, a zero correlation between the random intercept and random slope is assumed in this paper (i.e. $\sigma_{u01} = 0$). But even when $\sigma_{u01} \neq 0$ this independence restriction does not affect estimation and testing of the parameters if there are $\frac{1}{2}n_1$ pupils per treatment per class and if x_{ij} in the multilevel model in (2) is centered around zero (Raudenbush, 1993). The inclusion of random effects at each level of the multilevel data structure leads to the decomposition of the variance of a pupils THKS score y_{ij} into variance and covariance components, and correlated THKS scores of two pupils i and i' within the same class:

$$
\begin{aligned}
\mathrm{Var}(y_{ij}) &= \sigma^2_{u0} + \sigma^2_{u1} x^2_{ij} + \sigma^2_e, \text{ for each } i,j \text{ if } \sigma_{u01} = 0 \\
&= \sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_e \text{ since } x_{ij} = -1 \text{ or } +1 \\
\mathrm{Cov}(y_{ij}, y_{i'j}) &= \sigma^2_{u0} + x_{ij} x_{i'j} \sigma^2_{u1}, \text{ for each } i \neq i', \text{ if } \sigma_{u01} = 0 \\
&= \sigma^2_{u0} + \sigma^2_{u1} \text{ if } x_{ij} = x_{i'j} \text{ and } \sigma_{u01} = 0 \\
&= \sigma^2_{u0} - \sigma^2_{u1} \text{ if } x_{ij} \neq x_{i'j} \text{ and } \sigma_{u01} = 0.
\end{aligned}
\tag{3}
$$

So, the Ordinary Least Squares (OLS) estimator of β_1, which assumes independent THKS scores, should not be used, except if both σ^2_{u0} and σ^2_{u1} are equal to zero. The Generalized Least Squares (GLS) estimator of β_1 can be used if the variance and covariance components are known. In most cases, however, the variance components are unknown and Maximum Likelihood (ML, Hartley and Rao, 1967) or Restricted Maximum Likelihood (REML, Patterson and Thompson, 1971) estimation is required. See Searle, Casella and McCulloch (1992) for a description of these methods. It can be shown that in case of normally distributed outcomes ML and REML estimation correspond to Goldsteins Iterative Generalized Least Squares (IGLS, Goldstein, 1986) or Restricted Iterative Generalized Least Squares (RIGLS, Goldstein, 1989), respectively. Finally, the null hypothesis of no treatment effect may be tested using the test statistic $F = t^2 = \hat{\beta}^2_1 / \hat{\mathrm{Var}}(\hat{\beta}_1)$, which, under the null hypothesis, has an F distribution with 1 and n_2-1 degrees of freedom (Bryk and Raudenbush, 1992) when RIGLS is used. For models with a fixed slope β_1 (i.e. without interaction between class and treatment, $\sigma^2_{u1} = 0$), this test statistic has an F distribution with 1 and $n_1 n_2 - n_2 - 1$ degrees of freedom under the null hypothesis when RIGLS is used. The degrees of freedom are equal to those for the ANOVA models in the next section since RIGLS estimates are equal to REML estimates under the normal case (see Goldstein, 1989), and for balanced data the solutions of the REML equations are equal to ANOVA estimators, whether normality is assumed or not (see for more details Searle, Casella, and McCulloch, 1992, Sections 4.8 and 6.6f). Both IGLS and RIGLS are implemented in MLwiN (Goldstein *et al.*, 1998), a computer program for multilevel analysis, which we used for the analysis of the data sets in this paper. Multilevel analysis may also be done using the programs HLM (Bryk, Raudenbush, and Congdon, 1996), MIXREG (Hedeker and Gibbons, 1996), STATA (Stata Corporation, 2001), or the SAS (SAS Institute, 1996) module PROC MIXED.

3.1.2 Mixed effects ANOVA model

The multilevel model in (2) can also be expressed as a mixed effects ANOVA model. We have a factorial design where classes and treatment conditions are crossed for Design 1. There are two treatments, n_2 classes, and $\frac{1}{2}n_1$ pupils per treatment in each class. The ANOVA model for pupil i within class j and treatment t is given by

$$y_{ijt} = \mu + \alpha_t + u_j + (\alpha u)_{jt} + e_{ijt}, \tag{4}$$

where μ is the grand mean, α_t is the fixed effect associated with the t-th treatment, u_j is the random effect associated with the j-th class, $(\alpha u)_{jt}$ is the random interaction effect, and e_{ijt} is the random error term at the pupil level. If $t = 2$ for the treatment group and $t = 1$ for the control group, the correspondence between the parameters in the mixed effects ANOVA model (4) and the parameters in the multilevel regression model (2) is given by

$$\mu = \beta_0, \quad \frac{\alpha_2 - \alpha_1}{2} = \beta_1, \quad u_j = u_{0j}, \quad \frac{(\alpha u)_{j2} - (\alpha u)_{j1}}{2} = u_{1j}, \quad e_{ijt} = e_{ij}. \tag{5}$$

In principle there are two possibilities for the mixed effects ANOVA model: one without restrictions on the interaction terms, and one where $\Sigma_t (\alpha u)_{jt} = 0$ for all j (Searle *et al.*, 1992, section 4.3). This paper assumes the latter because this also applies to the multilevel model (since $\Sigma_i x_{ij} u_{1j} = 0$ for all j since there are only two treatment conditions which are coded -1 and +1, and there are $\frac{1}{2}n_1$ pupils per treatment per class). As in the multilevel model in (2), the random terms u_j, $(\alpha u)_{jt}$, and e_{ijt} are independently and normally distributed with zero mean and variances σ^2_{u0}, σ^2_{u1}, and σ^2_e, respectively. Note that the classical mixed effects ANOVA model assumes that the class effect and the treatment-by-class interaction are independent, which is not necessarily the case for multilevel models, where σ_{u01} may be unequal to zero. The variance components in ANOVA models are estimated by equating the Mean Squares (*MS*) to their expected values. In the lower half of Table 1, the expected Mean Squares $E(MS_{\text{mixed}})$ for the mixed effects factorial ANOVA model are given. The null hypothesis of no treatment effect can be tested using the test statistic $F = MS_{\text{treatment}} / MS_{\text{interaction}}$, which has an F distribution with 1 and n_2-1 degrees of freedom under the null hypothesis. With no interaction between class and treatment (i.e. $\sigma^2_{u1} = 0$), the Sum of Squares $SS_{\text{interaction}}$ is pooled with SS_{error}, and the test statistic becomes $F = MS_{\text{treatment}} / MS_{\text{error}}$, which has an F distribution with 1 and $n_1 n_2$-n_2-1 degrees of freedom under the null hypothesis. As will be shown in the following example, the mixed effects ANOVA model gives the same results as the multilevel regression model.

Example: Analyses of data set for Design 1

In this paper we will restrict ourselves to those Los Angeles schools which were randomized to either the media or no-treatment intervention group. Analysis of the data with two levels of nesting (pupils within classes) gave $\hat{\beta}_0 = 2.34$, $\hat{\beta}_1 = 0.12$, $\widehat{\sigma^2_{u0} + \sigma^2_{u1}} = 0.16$, $\hat{\sigma}^2_e = 1.72$. Note that the variance components cannot be estimated separately as there was only one treatment condition per class, see Section 3.2. For illustrative purposes we generated data for $n_2 = 70$ classes with $n_1 = 12$ pupils each, using the estimated regression coefficients and variance components from the real data set (with $\sigma^2_{u0} = 0.1$ and $\sigma^2_{u1} = 0.06$). These data were analyzed using the multilevel model and with REML estimation and using ANOVA estimation. The results of the analyses are presented in Table 1, showing that both approaches

give the same estimated regression coefficients and variance components, and both do not reject the null hypothesis of no treatment effect at the five percent level.

3.2 Design 2: Randomization at the class level

3.2.1 Multilevel regression model

The multilevel regression model for Design 2 is given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}, \tag{6}$$

where the random interaction term $x_{ij} u_{1j}$ in (2) is omitted since all pupils within the same class receive the same treatment. As a result σ^2_{u0} and σ^2_{u1} cannot be estimated separately. Instead, their sum is estimated which will be coded as $\sigma^2_u = \sigma^2_{u0} + \sigma^2_{u1}$ in this paper. Furthermore x_{ij} may be replaced by x_j since treatment condition does not vary within classes, and again treatment condition is coded as $x_j = +1$ for the treatment group and $x_j = -1$ for the control

Table 1 Results for Data Set for Design 1

Results based on the multilevel model

Parameter	Estimate	Standard Error	$F = t^2$	p
Fixed effects:				
Intercept, β_0	2.304	0.058		
Treatment effect, β_1	0.097	0.050	3.776	0.056
Random effects:				
Random intercept, σ^2_{u0}	0.096			
Random slope, σ^2_{u1}	0.039			
Random error term, σ^2_e	1.634			

Results based on the mixed effects ANOVA model

Source	df		SS	$MS = SS/df$	E(MS_{mixed})	F	p
Treatment	1	= 1	7.947	7.947	$\sigma^2_e + n_1\sigma^2_{u1} + \frac{1}{2}n_1 n_2 \sigma^2_\alpha$	3.776	0.056
Class	n_2-1	= 69	192.148	2.785	$\sigma^2_e + n_1\sigma^2_{u0}$		
Interaction	n_2-1	= 69	145.224	2.105	$\sigma^2_e + n_1\sigma^2_{u1}$		
Error	$n_1 n_2$-2n_2	= 700	1144.088	1.634	σ^2_e		
Total	$n_1 n_2$-1	= 839	1489.407				

$\hat{\sigma}^2_{u0} = (MS_{\text{class}} - MS_{\text{error}})/12 = 0.096$, $\hat{\sigma}^2_{u1} = (MS_{\text{interaction}} - MS_{\text{error}})/12 = 0.039$, $\hat{\sigma}^2_e = MS_{\text{error}} = 1.634$

$\hat{\mu} = \bar{y}_{...} = 2.304$, $(\hat{\alpha}_2 - \hat{\alpha}_1)/2 = (\bar{y}_{.2} - \bar{y}_{.1})/2 = 0.097$

Note. It is assumed that $\sum_t (\alpha u)_{jt} = 0$. Without this restriction n_1 has to be replaced by $\frac{1}{2}n_1$ in the E(MS) for treatment and interaction (Searle *et al.*, 1992, p. 123-126).

group. The null hypothesis of no treatment effect is tested by the test statistic $F = t^2 = \hat{\beta}_1^2/\widehat{\text{Var}}(\hat{\beta}_1)$ which, under the null hypothesis, has an F distribution with 1 and n_2-2 degrees of freedom (Bryk and Raudenbush, 1992) when RIGLS is used. The degrees of freedom are equal to those for the ANOVA model in the next section (see Section 2.3.1.1 for the explanation). Note that the degrees of freedom depend on the level of randomization and the presence of treatment by class interaction.

3.2.2 Mixed effects ANOVA model

For randomization at the class level, we have a mixed effects nested ANOVA model in which classes are nested within treatments:

$$y_{ijt} = \mu + \alpha_t + u_{jt} + e_{ijt}, \tag{7}$$

where μ is the grand mean, α_t is the fixed effect associated with the t-th treatment, and u_{jt} and e_{ij} are the random terms at the class and pupil level which are assumed to be independently and normally distributed with zero mean and variances σ_u^2 and σ_e^2, respectively. There are $\frac{1}{2}n_2$ classes per treatment and n_1 pupils per class. If $t = 2$ for the treatment group and $t = 1$ for the control group, the correspondence between the mixed effects ANOVA model in (7) and the multilevel regression model in (6) is given by

$$\mu = \beta_0, \qquad \frac{\alpha_2 - \alpha_1}{2} = \beta_1, \qquad u_{jt} = u_j, \qquad e_{ijt} = e_{ij}. \tag{8}$$

The expected Mean Squares for the ANOVA model in (7), $E(MS_{\text{mixed}})$, are given in the lower half of Table 2. The test statistic for the null hypothesis of no treatment effect is given by $F = MS_{\text{treatment}} / MS_{\text{class}}$ which, under the null hypothesis, has an F distribution with 1 and n_2-2 degrees of freedom.

Example: Analyses of data set for Design 2

A data set was also generated for randomization at the class level with $n_1 = 12$ and $n_2 = 70$, using the estimated parameter values from the analysis of the real data as input for the generation process. The data were analyzed using the multilevel model with REML estimation and using ANOVA estimation. Again both methods produce the same estimated parameter values and test statistic and both reject the null hypothesis of no treatment effect at the five percent level, see Table 2 for the numerical results. Note that Design 1 shows a non-significant effect, while Design 2 shows a significant effect, which is a consequence of the fact that the generated effect for Design 2 is larger than the generated effect for Design 1.

Table 2 Results for Data Set for Design 2

Results based on the multilevel model

Parameter	Estimate	Standard Error	$F = t^2$	p
Fixed effects:				
Intercept, β_0	2.256	0.073		
Treatment effect, β_1	0.166	0.073	5.221	0.025
Random effects:				
Random intercept, σ_u^2	0.212			
Random error term, σ_e^2	1.905			

Results based on the mixed effects ANOVA model

Source	df		SS	$MS = SS/df$	$E(MS_{mixed})$	F	p
Treatment	1	= 1	23.231	23.231	$\sigma_e^2 + n_1\sigma_u^2 + \frac{1}{2}n_1n_2\sigma_\alpha^2$	5.221	0.025
Classes within treatments	n_2-2	= 68	302.599	4.450	$\sigma_e^2 + n_1\sigma_u^2$		
Error	$n_1n_2 - n_2$	= 770	1466.541	1.905	σ_e^2		
Total	n_1n_2-1	= 839	1792.371				

$\hat{\sigma}_u^2 = (MS_{class} - MS_{error})/12 = 0.212$, $\hat{\sigma}_e^2 = MS_{error} = 1.905$.

$\hat{\mu} = \bar{y}_{...} = 2.256$, $(\hat{\alpha}_2 - \hat{\alpha}_1)/2 = (\bar{y}_{.2} - \bar{y}_{.1})/2 = 0.166$

4 Traditional models

Three more traditional regression models for the analysis of multilevel experimental data are the fixed effects model, the disaggregated data model and the aggregated data model. These models are presented in this section, together with their equivalent ANOVA models.

4.1 Fixed effects regression model and fixed effects ANOVA model

In contrast to the multilevel regression model, the fixed effects regression model includes class and interaction effects as fixed effects. For randomization at the pupil level (Design 1) we have

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{h=1}^{n_2-1} \beta_{h+1} d_{hj} + \sum_{h=1}^{n_2-1} \beta_{n_2+h} x_{ij} d_{hj} + e_{ij}, \tag{9}$$

where y_{ij} is the THKS score, x_{ij} denotes treatment condition, and e_{ij} is the random error term at the pupil level with zero mean and variance σ_e^2. The classes may be represented by n_2-1 dummy variables d_h and the n_2-th class is the reference class. The dummy variables are coded such that $d_{hj} = +1$ if $h = j$, $d_{hj} = -1$ if $h = n_2$, and $d_{hj} = 0$ otherwise, so that they are centered and $x_{ij}\, d_{hj}$ is orthogonal to x_{ij} and d_{hj}. The fixed effects ANOVA model is given by (4), where u_j is

now regarded a fixed effect. The corresponding expected mean squares are equal to those for the mixed effects ANOVA model in Table 1 except that $E(MS_{\text{treatment}}) = \sigma_e^2 + \frac{1}{2}n_1 n_2 \sigma_\alpha^2$. The test statistic for the null hypothesis of no treatment effect is calculated as $F = \hat{\beta}_1^2/\hat{\text{Var}}(\hat{\beta}_1) = MS_{\text{treatment}} / MS_{\text{error}}$ and has 1 and $n_1 n_2 - 2n_2$ degrees of freedom under the null hypothesis. The terms $x_{ij}d_{hj}$ may be deleted from model (9) if there is no interaction between treatment and class, and the $SS_{\text{interaction}}$ is pooled with the SS_{error}. Then, the null hypothesis is tested using the statistic $F = \hat{\beta}_1^2/\hat{\text{Var}}(\hat{\beta}_1) = MS_{\text{treatment}} / MS_{\text{error}}$, which has an F distribution with 1 and $n_1 n_2 - n_2 - 1$ degrees of freedom under the null hypothesis.

There are one reference class and $\frac{1}{2}n_2 - 1$ centered dummy variables for each treatment condition and no interaction terms in the fixed effect regression model in (9) for randomization at the class level. The fixed effects ANOVA model is given by (7), with u_{jt} being a fixed effect. The corresponding expected mean squares are equal to those for the mixed effects ANOVA model in Table 2 except that $E(MS_{\text{treatment}}) = \sigma_e^2 + \frac{1}{2}n_1 n_2 \sigma_\alpha^2$. The test statistic for the null hypothesis is calculated as $F = \hat{\beta}_1^2/\hat{\text{Var}}(\hat{\beta}_1) = MS_{\text{treatment}} / MS_{\text{error}}$, which has 1 and $n_1 n_2 - n_2$ degrees of freedom under the null hypothesis.

4.2 Disaggregated data model and one-way ANOVA model

The disaggregated data model differs from the multilevel model in (2) in that it contains only one random error term:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + r_{ij}, \tag{10}$$

where y_{ij} and x_{ij} are the THKS score and treatment condition, respectively, and r_{ij} is a random error term with zero mean and variance σ_r^2. Note that $r_{ij} = u_{0j} + u_{1j}x_{ij} + e_{ij}$ if the mixed effect model is correct but the disaggregated data model is assumed. In that case the variance of a pupil's THKS score is equal to $\text{Var}(y_{ij}) = \sigma_r^2 = \sigma_{u0}^2 + \sigma_{u1}^2 + \sigma_e^2$, according to (3). However, the disaggregated data model assumes that the r_{ij}'s are independent, that is: $\text{Cov}(y_{ij}, y_{i'j}) = 0$, for $i \neq i'$, and so the disaggregated data model ignores dependence of THKS scores of pupils within the same class and the OLS estimator can be used in this case. Like in the multilevel modelling approach, the test statistic for the test of no treatment effect is given by $F = t^2 = \hat{\beta}_1^2/\hat{\text{Var}}(\hat{\beta}_1)$, which follows an F distribution with 1 and $n_1 n_2 - 2$ degrees of freedom under the null hypothesis of no treatment effect.

The ANOVA model corresponding to the disaggregated data model in (10) is given by

$$y_{it} = \mu + \alpha_t + r_{it}, \tag{11}$$

where y_{it} is the THKS score for the i-th pupil within the t-th treatment, μ is the grand mean, α_t is the fixed effect for the t-th treatment. The random error term at the pupil level r_{it} has variance σ_r^2. There are $\frac{1}{2}n_1 n_2$ pupils per treatment. The null hypothesis of no treatment effect is now tested using the test statistic $F = MS_{\text{treatment}} / MS_{\text{error}}$, which has an F distribution with 1 and $n_1 n_2 - 2$ degrees of freedom under the null hypothesis. Note that the disaggregated data model in (10) and the one-way ANOVA model in (11) both apply to Design 1 and Design 2 since they ignore class effects altogether.

4.3 Aggregated data model

The aggregated data model describes the data after aggregation (averaging) within the same treatment to the class level. This means that each cell in Figure 1 collapses to one average observation. Thus, if randomization is done at the pupil level, we have two correlated mean THKS scores per class, one for each treatment condition. Based upon (2) the regression models for the mean outcomes in the control and treatment group, denoted by $\bar{y}_{.jc}$ and $\bar{y}_{.jt}$, are given by

$$\begin{aligned} \bar{y}_{.jc} &= \beta_0 - \beta_1 + u_{0j} - u_{1j} + \bar{e}_{.jc} \\ \bar{y}_{.jt} &= \beta_0 + \beta_1 + u_{0j} + u_{1j} + \bar{e}_{.jt}, \end{aligned} \tag{12}$$

respectively. The treatment effect per class can be estimated by

$$\hat{\beta}_{1j} = \frac{\bar{y}_{.jt} - \bar{y}_{.jc}}{2}, \tag{13}$$

which has variance $\sigma_{u1}^2 + \sigma_e^2/n_1$. For equal class sizes the overall treatment effect is then estimated as the mean of the $\hat{\beta}_{1j}$, which has variance $(n_1\sigma_{u1}^2 + \sigma_e^2)/n_1 n_2$. To test the null hypothesis of no treatment effect the paired samples t-test with n_2-1 degrees of freedom under the null hypothesis can be used.

If randomization is done at the class level, there is one mean THKS score per class, $\bar{y}_{.j}$, which is related to treatment condition x_j, by

$$\bar{y}_{.j} = \beta_0 + \beta_1 x_j + u_j + \bar{e}_{.j}, \tag{14}$$

where $\bar{e}_{.j}$ is the class mean of the random effect at the pupil level. The mean THKS scores per class are assumed to be independently distributed with variance

$$\mathrm{Var}(\bar{y}_{.j}) = \mathrm{Var}(u_j + \bar{e}_{.j}) = \sigma_u^2 + \frac{\sigma_e^2}{n_1}, \tag{15}$$

and the OLS estimator can be used for the aggregated data model. The null hypothesis of no treatment effect can be tested with the independent samples t-test with n_2-2 degrees of freedom under the null hypothesis.

5 Analyses of artificial data sets

The two data sets from Section 3 were analyzed with the multilevel model, the fixed effects model, the disaggregated data model, and the aggregated data model. REML estimation (using the computer program MLwiN, Goldstein *et al.*, 1998) was used for the multilevel model, and OLS estimation (using SPSS, SPSS Inc, 1998) for all other models.

The results for Design 1 are given in the upper part of Table 3. This table shows that all models produce the same estimated treatment effect $\hat{\beta}_1$, but that its standard error is

Table 3 Results of Multilevel and Traditional Analyses of Data Sets

	Model			
	Multilevel	Fixed effects	Disaggregated data	Aggregated data
		Design 1: Randomization pupil level		
$\hat{\sigma}^2_{u0}$	0.096	-	-	-
$\hat{\sigma}^2_{u1}$	0.039	-	-	-
$\hat{\sigma}^2_e$	1.634	1.634	-	-
$\hat{\sigma}^2_r$	-	-	1.768	-
$\frac{\hat{\sigma}^2_e}{12}+\hat{\sigma}^2_{u1}$	-	-	-	0.175
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.097 (0.050)	0.097 (0.044)	0.097 (0.046)	0.097 (0.050)
t_{β_1} (*df*)	1.943 (69)	2.205 (700)	2.120 (838)	1.943 (69)
p-value	0.056	0.028	0.034	0.056
		Design 2: Randomization class level		
$\hat{\sigma}^2_u$	0.212	-	-	-
$\hat{\sigma}^2_e$	1.905	1.905	-	-
$\hat{\sigma}^2_r$	-	-	2.111	-
$\frac{\hat{\sigma}^2_e}{12}+\hat{\sigma}^2_u$	-	-	-	0.371
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.166 (0.073)	0.166 (0.048)	0.166 (0.050)	0.166 (0.073)
t_{β_1} (*df*)	2.285 (68)	3.492 (770)	3.317 (838)	2.284 (68)
p-value	0.025	0.001	0.001	0.025

underestimated by the disaggregated data model and the fixed effects model. As a result, the test statistics for these two models are somewhat larger than those for the multilevel model and the aggregated data model, and *p*-values for the fixed effects model and the disaggregated data model are too small. For these two models the null hypothesis of no treatment effect is rejected at the 5% level, which was not the case for the multilevel model and aggregated data model. Thus, the use of the fixed effect model and disaggregated data model leads to too liberal statistical tests on the treatment effect. Note that for the aggregated data model σ^2_e cannot be disentangled from σ^2_{u1} since there are only two mean THKS scores within each class. Note also, however, that the aggregated data model yields the same results as the multilevel analysis.

The results of the analysis for Design 2 are given in the lower part of Table 3, showing that, again, the multilevel model and aggregated data model produce the same estimated treatment effect and standard error, whereas the latter is too small for the fixed effects model and the disaggregated data model. All models reject the null hypothesis of no treatment effect at the five percent level, but this is not necessarily the case for other data sets. Note that σ_e^2 cannot be disentangled from σ_u^2 for the aggregated data model since there is only one mean THKS score per class.

6 Comparison of the four methods based on analytical expressions

In this section the results in the previous section will be explained by means of analytical expressions. The four estimation methods will be compared with each other assuming non-varying class sizes and no covariates. When control and treatment groups are coded by $x_{ij} = -1$ and $x_{ij} = +1$, the estimator of the treatment effect β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{ij} x_{ij} y_{ij}}{n_1 n_2} = \frac{\bar{y}_{..t} - \bar{y}_{..c}}{2}, \tag{16}$$

for each of the four models and for both levels of randomization. The means $\bar{y}_{..c}$ and $\bar{y}_{..t}$ are the mean THKS scores in the control and treatment groups, respectively.

Table 4. $\hat{\mathrm{Var}}(\hat{\beta}_1)$ for the four Regression Models

	Model			
Level of randomization	Multilevel	Fixed effects	Disaggregated data	Aggregated data
Pupil (interaction treatment by class)	$\frac{n_1\hat{\sigma}_{u1}^2+\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_r^2}{n_1 n_2} = \frac{\sigma_{u0}^2+\sigma_{u1}^2+\sigma_e^2}{n_1 n_2}$	$\frac{n_1\sigma_{u1}^2+\sigma_e^2}{n_1 n_2}$
Pupil (no interaction treatment by class)	$\frac{\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_r^2}{n_1 n_2} = \frac{\sigma_{u0}^2+\sigma_e^2}{n_1 n_2}$	$\frac{n_1\sigma_{u1}^2+\sigma_e^2}{n_1 n_2}$
Class	$\frac{n_1\hat{\sigma}_u^2+\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_e^2}{n_1 n_2}$	$\frac{\hat{\sigma}_r^2}{n_1 n_2} = \frac{\sigma_u^2+\sigma_e^2}{n_1 n_2}$	$\frac{n_1\sigma_u^2+\sigma_e^2}{n_1 n_2}$

Note. Control and treatment group are denoted by $x_{ij} = -1$ and $x_{ij} = +1$, respectively.
For class level randomization $\sigma_u^2 = \sigma_{u0}^2 + \sigma_{u1}^2$.
For the aggregated data model the variance components cannot be estimated separately. Furthermore, for randomization at the pupil level the aggregated data model always assumes treatment by class interaction.

In Table 4 the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ for the four models are given for both class and pupil level randomization. The second column gives the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ that is obtained when the multilevel model is applied to the data. This model uses random effects to represent the classes in the study so that the results from the study may be generalized to the whole population from which the classes were sampled. The fixed effects model, on the other hand, uses fixed effects to represent the classes. The $\hat{\mathrm{Var}}(\hat{\beta}_1)$ obtained with this model is given in the third column and does not depend on the level of randomization or on the presence or absence of treatment by class interaction. Since the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ obtained with the fixed effects model may be underestimated compared with that obtained with the multilevel model, the fixed effects model should not be used as an alternative to the multilevel model at least if we want to consider the clusters in our study as a random sample from some population to which we want to generalize our results. The amount of underestimation obtained with the fixed effects model can be quantified with the ratio of the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ obtained with both models which depends on n_1 and the intra-class correlation coefficient ρ_c, which measures the amount of variance between classes, i.e. $\rho_c=\sigma_u^2/(\sigma_u^2+\sigma_e^2)$. For example, for cluster level randomization (Design 2) this variance ratio is equal to

$$\text{variance ratio} = \frac{\hat{\mathrm{Var}}(\hat{\beta}_1)_{\text{multilevel}}}{\hat{\mathrm{Var}}(\hat{\beta}_1)_{\text{fixed effects}}} = \frac{n_1\hat{\sigma}_u^2+\hat{\sigma}_e^2}{\hat{\sigma}_e^2} = n_1\left(\frac{\hat{\rho}_c}{1-\hat{\rho}_c}\right)+1, \tag{17}$$

and increases with both ρ_c and n_1. Even for small intra-class correlation coefficient the amount of underestimation may be unacceptable. For example, if $\rho_c = 0.1$ and $n_1 = 30$ the variance ratio is approximately equal to 4. So, the confidence interval for β_1 in the fixed effects model will be about twice as small as that obtained under the multilevel model, and the null hypothesis of no treatment effect will be rejected too often, leading to an inflation of the type I error rate. Thus the fixed effects model may result in incorrect conclusions, which may also be the case for other values of ρ_c and n_1 and for pupil level randomization (Design 1). Consequently it should not be used for the analysis of multilevel data when generalizations have to be made to an underlying population of classes.

The fourth column in Table 4 gives the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ that is obtained when the nesting of pupils within classes is ignored, thus when the disaggregated data model is used. For class level randomization the $\hat{\mathrm{Var}}(\hat{\beta}_1)$ is underestimated. The amount of underestimation is given by the variance ratio

$$\text{variance ratio} = \frac{\hat{\mathrm{Var}}(\hat{\beta}_1)_{\text{multilevel}}}{\hat{\mathrm{Var}}(\hat{\beta}_1)_{\text{disaggregated data}}} = \frac{n_1\hat{\sigma}_u^2+\hat{\sigma}_e^2}{\hat{\sigma}_r^2} = n_1\hat{\rho}_c+(1-\hat{\rho}_c) = (n_1-1)\hat{\rho}_c+1. \tag{18}$$

which increases with both ρ_c and n_1. For example, if $\rho_c = 0.05$ and $n_1 = 30$, this ratio is equal to 2.45, and thus the confidence interval for β_1 obtained with the disaggregated data model is $\sqrt{2.45}=1.6$ times as small as that obtained with the multilevel model, and such a high value is not acceptable. For randomization at the pupil level and no treatment by class interaction the

$\hat{\mathbf{Var}}(\hat{\beta}_1)$ is slightly overestimated with the disaggregated data model. If treatment by class interaction is present the $\hat{\mathbf{Var}}(\hat{\beta}_1)$ is under- or overestimated by the disaggregated data model, depending on the values of the variance components and the class size. Thus, the disaggregated data model should not be used as an alternative to the multilevel model.

The last column of Table 4 gives the $\hat{\mathbf{Var}}(\hat{\beta}_1)$ that is obtained with the aggregated data model. For pupil level randomization and no treatment by cluster interaction the results for the aggregated data model (i.e. the paired samples *t*-test) are inefficient. In fact the aggregated data model assumes interaction between treatment and class since it calculates the test statistic as $F = t^2 = MS_{\text{treatment}} / MS_{\text{interaction}}$, thus the $\hat{\mathbf{Var}}(\hat{\beta}_1)$ is inefficiently but unbiasedly estimated and the degrees of freedom for the denominator for the test statistic are too low. For class level randomization and for pupil level randomization with treatment by cluster interaction this model yields the same $\hat{\mathbf{Var}}(\hat{\beta}_1)$ as the multilevel model. However, we do not in general recommend the aggregated data model to be used as an alternative to the multilevel model since in general class sizes vary which makes the use of the aggregated data model complicated as we will see in the next section.

The conclusions in this section are presented schematically in Table 5, which also gives references where some of the conclusions have also been presented.

Table 5. Comparison of Traditional Models to Multilevel Model with Respect to Estimated Var($\hat{\beta}_1$) and Degrees of Freedom of the Denominator for the *F* Distribution of the Test Statistic under the Null Hypothesis, Assuming Equal Class Sizes and Classes Represent a Random Sample

		Analysis model	
Level of randomization	Fixed effects	Disaggregated data	Aggregated data
Pupil (interaction treatment by class)	Underestimated Var($\hat{\beta}_1$), Senn (1998), Gould (1998), Jones *et al.* (1998)	Underestimated or overestimated Var($\hat{\beta}_1$), depending on values variance components.	Correctly estimated Var($\hat{\beta}_1$). Equal to paired samples *t*-test on class by treatment means.
	incorrect df: n_1n_2-$2n_2$	incorrect df: n_1n_2-2	correct df: n_2-1
Pupil (no interaction treatment by class)	Correctly estimated Var($\hat{\beta}_1$).	Overestimated Var($\hat{\beta}_1$), Parzen *et al.*(1998)	Inefficiently estimated Var($\hat{\beta}_1$) because equal to paired samples *t*-test on class by treatment means which assumes interaction.
	correct df: n_1n_2-n_2-1	incorrect df: n_1n_2-2	unnecessarily low df: n_2-1
Class	Underestimated Var($\hat{\beta}_1$).	Underestimated Var($\hat{\beta}_1$), Hedeker *et al.* (1994), Longford (1995), Barcikowski (1981)	Correctly estimated Var($\hat{\beta}_1$), Hopkins (1982). Equal to independent samples *t*-test on class means.
	incorrect df: n_1n_2-n_2	incorrect df: n_1n_2-2	correct df: n_2-2

7 Generalization to more complex models

The results in the previous section are limited to equal class sizes and models with no covariates. Equal class sizes may not always be feasible in practice, and in some cases covariates have to be included into the model. In this section, these restrictions will be relaxed one at a time.

7.1 Varying class sizes

In this section we will assume varying class sizes n_{1j}, but 50:50 randomization to the treatment and control group. So, there are $\frac{1}{2}n_{1j}$ pupils per treatment in class j for randomization at the pupil level, and $\frac{1}{2}n_2$ classes per treatment for randomization at the class level. The fixed effects model and the disaggregated data model differ from the multilevel model with respect to $\widehat{\mathbf{Var}}(\hat{\beta}_1)$ but not necessarily with respect to $\hat{\beta}_1$ itself. The results for the aggregated data model correspond to those of the multilevel model if weighting of class j is done by the factor $w_R = (\sigma_e^2/n_{1j} + \sigma_u^2)^{-1}/C$ for Design 2, where $\mathbf{C}=\sum_{k:x_k=x_j}(\sigma_e^2/n_{1k}+\sigma_u^2)^{-1}$, and $w_R = (\sigma_e^2/n_{1j} + \sigma_{u1}^2)^{-1}/C$ for Design 1 and assuming treatment by class interaction, where $\mathbf{C}=\sum_j(\sigma_e^2/n_{1j}+\sigma_{u1}^2)^{-1}$. However data aggregation with these weights is hardly a simple alternative to multilevel analysis if the variance components are unknown and have to be estimated. When σ_e^2/n_{1j} is large compared with σ_u^2 or σ_{u1}^2, these weights are almost equal to $w_{II} = n_{1j}/\Sigma_j n_{1j}$ and weighting is done according to the number of pupils per class. On the other hand if σ_e^2/n_{1j} is small these weights are almost equal to $w_{III} = 1/n_2$ which implies that no weighting is done. The treatment effect estimator with weighting by w_R will be bounded by the estimated treatment effects with weighting by w_{II} and w_{III}. It can be shown (Bloch and Moses, 1988) that an unweighted analysis is at most 12.5% less efficient than weighting by the proper weights w_R if $\sigma_u^2 \geq \sigma_e^2((\min n_{1j})^{-1}-2(\max n_{1j})^{-1})$ within each treatment condition for class level randomization, or $\sigma_{u1}^2 \geq \sigma_e^2((\min n_{1j})^{-1}-2(\max n_{1j})^{-1})$ overall for randomization at the pupil level. A sufficient but not necessary condition is $(\max n_{1j})/(\min n_{1j}) \leq 2$ overall for pupil level randomization, or within each treatment condition for class level randomization. On the other hand, the weights w_{II} are generally preferred when treating classes as fixed (Lin, 1999).

Example: Analysis of TVSFP data

To compare the traditional models with the multilevel model in the case of varying class sizes the TVSFP data, with restriction to the Los Angeles pupils in the media or no-treatment control group, were analyzed. In the analyses two levels of nesting are taken into account: pupils within classes. Class sizes ranged from 1 till 27 with a mean of 12 pupils per class. All pupils within a class received the same treatment condition and the interaction between treatment condition and class cannot be estimated. Treatment condition was used as the only explanatory variable to model the outcome THKS, leaving the inclusion of the pre-treatment THKS to Section 7.2.

The results of the analyses are presented in the upper half of Table 6. Compared with

the multilevel model, the fixed effects model and the disaggregated data model both produce too large estimates of the treatment effect and too small standard errors. As a result test statistics are too large and *p*-values are too small, which was also true for non-varying class sizes (see Section 6). Furthermore the estimated treatment effect of the multilevel model is bounded by those of the aggregated data models with weighting according to cluster size and without weighting. The latter even produces an estimated treatment effect below zero which in this case is a result of not taking class sizes into account. Weighting according to class size results in a standard error which is smaller than that for the multilevel model due to the fact that this type of weighting ignores intra class correlation. The estimated treatment effect of the fixed effects model corresponds to that of the disaggregated data model and the aggregated data model with weighting according to class size since the dummy variables of the fixed effects model are coded such that they are orthogonal to treatment x_j.

Table 6. Results of multilevel and traditional analyses of TVSFP data

	Model				
	Multilevel	Fixed effects	Disaggregated data	Aggregated data weighting by w_{II}	Aggregated data no weighting(w_{III})
Model without pre-test THKS					
$\hat{\sigma}_u^2$	0.166	-	-	-	-
$\hat{\sigma}_e^2$	1.718	1.708	-	-	-
$\hat{\sigma}_r^2$	-	-	1.871	-	-
$\hat{\beta}_1$ ($\hat{SE}(\hat{\beta}_1)$)	0.056 (0.070)	0.089 (0.045)	0.089 (0.047)	0.089 (0.067)	-0.041 (0.082)
$t_{\beta_1}(df)$	0.8011 (68)	1.964 (767)	1.876 (835)	1.331 (68)	-0.498 (68)
p-value	0.426	0.050	0.061	0.188	0.620
Model with pre-test THKS					
$\hat{\sigma}_u^2$	0.107	-	-	-	-
$\hat{\sigma}_e^2$	1.573	1.557	-	-	-
$\hat{\sigma}_r^2$	-	-	1.674	-	-
$\hat{\beta}_1$ ($\hat{SE}(\hat{\beta}_1)$)	0.085 (0.061)	0.138 (0.046)	0.106 (0.045)	0.106 (0.060)	-0.018 (0.079)
$t_{\beta_1}(df)$	1.340 (66)	2.966 (766)	2.369 (833)	1.727 (67)	-0.229 (67)
p-value	0.168	0.003	0.018	0.082	0.819

7.2 Models with covariates

A covariate c_{ij} may be split into a component $\bar{c}_j$ which varies only between classes, and a component c_{ij} - $\bar{c}_j$ which varies only within classes (Neuhaus and Kalbfleisch, 1998). These two components may then be added to the multilevel model (2), the fixed effects model (9), the disaggregated data model (10), and the aggregated data models (12) and (14). Then it can be shown that the $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ given in Table 6 have to be divided by $(1\text{-}R^2)$, where R^2 is the squared multiple correlation between treatment condition and the other independent variables in the model, i.e. the two components of the covariate, and for the fixed effects model the dummy variables as well.

For randomization at the class level $(1\text{-}R^2)$ is equal to $(1-r^2_{x_j,\bar{c}_j})$ for the multilevel model, the disaggregated data model, and the aggregated data model, and the comparison for these models made in the previous section will roughly hold. Since there is multicollinearity between treatment condition, the n_2-2 dummy variables, and the class component of the covariate $\bar{c}_j$, the latter cannot be added to the fixed effects model. As the dummy variables (which have been centered) and the pupil component of the covariate c_{ij} - $\bar{c}_j$ are orthogonal to the treatment effect, $(1\text{-}R^2)$ is equal to 1 for the fixed effects model.

For randomization at the pupil level and assuming no treatment by class interaction, the $(1\text{-}R^2)$ equals $(1-r^2_{x_{ij},c_{ij}-\bar{c}_j})$ for the multilevel model and the disaggregated data model, and the comparison for these two models as made in Section 5 will still hold. Since there is multicollinearity between the n_2-1 dummy variables, and the class component of the covariate $\bar{c}_j$, the latter cannot be added to the fixed effects model. As the dummy variables are orthogonal to the treatment effect, $(1\text{-}R^2)$ is the same as for the multilevel model. The pupil level component c_{ij} - $\bar{c}_j$ of the covariate is equal to zero in the aggregated data model, so the pupil level variance for the aggregated data model will be larger than for the other models, but also will the term $(1\text{-}R^2)$ be equal to zero for the aggregated data model. Assuming treatment by class interaction, however, the formula for the $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ in the presence of covariates will become more complex and are beyond the scope of this paper.

Example: Analysis of TVSFP data (continued)

The pre-treatment THKS was split into a component which varies at the class level and one which varies at the pupil level, and both components were added to the model as covariates. As a result, the estimated variance components at both levels will decrease. The results of the analyses are given in the lower part of Table 6. Observed *p*-values were too low for the fixed effects model, the disaggregated data model and the aggregated data model with weighting according to class size, leading to an incorrect rejection of the null hypothesis. Note that the treatment effect estimate according to the fixed effects model differs from the estimate by the disaggregated data model although the dummy variables are orthogonal to the treatment factor. This is due to the fact that both dummy variables and treatment factor slightly correlate with the class level covariate. The estimated treatment effect of the disaggregated data model corresponds to that of the aggregated data model with weighting according to class size, whereas the estimated treatment effect for the multilevel model is bounded by those of the aggregated data model with and without weighting.

8 Discussion and conclusions

In this study four regression models for the analysis of multilevel experimental data were compared: the multilevel model, the fixed effects model, the disaggregated data model, and the aggregated data model. To show the similarities with familiar ANOVA models, these models were also presented in terms of ANOVA notation. It was assumed that the conditions for random sampling of clusters from a larger population of clusters were satisfied, so that the experimental results were not only valid for the clusters involved in the study, but could also be generalized to the population of clusters. In that case the multilevel model should be used for the data analysis, but as this model is relatively new and rather complex, it was investigated whether the fixed effects model, the disaggregated data model, and the aggregated data model could be used as an alternative to the multilevel model. As criterion for the comparison the estimator of the treatment effect $\hat{\beta}_1$ and its variance $\mathbf{V\hat{a}r}(\hat{\beta}_1)$ were used, since these are generally of main interest in such experimental evaluations of treatments.

The results of the analyses of simulated and real data, and the analytical formulae for $\hat{\beta}_1$ and $\mathbf{V\hat{a}r}(\hat{\beta}_1)$ show that the use of the fixed effects model, and the disaggregated data model may result in incorrect estimates of the treatment effect and its standard error. Consequently these two models may yield conclusions on the treatment effect that differ from those obtained with the multilevel model. For varying cluster sizes data aggregation without weighting is less efficient than multilevel analysis. In order to calculate the correct weights for an aggregated data analysis the values of the variance components needs to be known. Furthermore, the use of the aggregated data model leads to a loss of information when the model contains covariates. Therefore, the multilevel model is the only model that may be used when the study results have to be generalized to the whole underlying population of clusters from which the clusters in the study are assumed to represent a random sample. The parameters in the multilevel model should be estimated using maximum likelihood or restricted maximum likelihood estimation, for instance with the computer programs mentioned in Section 3.1.1 which are especially designed for multilevel data.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society, Series A*, **149**, 1-43.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis, *Journal of Educational Statistics*, **6**, 267-285.

Bass, M. J., McWinney, I. R., & Donner, A. (1986). Do family physicians need medical assistance to detect and manage hypertension?, *Canadian Medical Association Journal*, **134**, 1247-1255.

Bloch, D. A., & Moses, L. E. (1988). Nonoptimally weighted least squares, *The American Statistician*, **42**, 50-53.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park: Sage

Publications.

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

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling, *Statistics in Medicine*, **17**, 1261-1291.

Dunlop, D. D. (1994). Regression for longitudinal data: A bridge from least squares regression, *The American Statistician*, **48**, 299-303.

Flay, B. R., Brannon, B. R., Johnson, C. A., Hansen, W. B., Ulene, A. L., Whitney- Santiel, D. A., Gleason, L. R., Sussman, S., Gavin, M. D., Glowacz, K. M., Sobol, D. F., & Spiegel, D. C. (1988). The television school and family smoking prevention and cessation project. I. Theoretical basis and program development, *Preventive Medicine*, **17**, 585-607.

Flay, B. R., Miller, T. Q., Hedeker, D., Siddiqui, O., Britton, C. F., Brannon, B. R., Johnson, C. A., Hansen, W. B., Sussman, S., & Dent C. (1995). The television, school, and family smoking prevention and cessation project. VIII. Student outcomes and mediating variables, *Preventive Medicine*, **24**, 29-40.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares, *Biometrika*, **73**, 43-56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least squares estimation, *Biometrika*, **76**, 622-623.

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

Goldstein, H, Rasbash, J, Plewis, I, Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. London: Institute of Education.

Gould, A. L. (1998). Multi-centre trial analysis revisited, *Statistics in Medicine*, **17**, 1779-1797

Hartley, H. O., & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika*, **54**, 93-108.

Hedeker, D., Gibbons, R. D., & Davis, J. M. (1991). Random regression models for multicenter clinical trials data, *Psychopharmacology Bulletin*, **27**, 73-77.

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

Hedeker, D., McMahon, S. D., Jason, L. A., & Salina, D. (1994). Analysis of clustered data in community psychology: With an example from a worksite smoking cessation project, *American Journal of Community Psychology*, **22**, 595-615.

Hopkins, K. D. (1982). The unit of analysis: group means versus individual observations, *American Educational Research Journal*, **19**, 5-18 .

Hox, J. J. (1994). *Applied multilevel analysis*. Amsterdam: TT-Publikaties

Jones, B., Teather, J. W., & Lewis, J. A. (1998). A comparison of various estimator of

treatment difference for a multi-centre clinical trial, *Statistics in Medicine*, **17**, 1767-1777.

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

Lin, Z. (1999). An issue of statistical analysis in controlled multi-centre studies: how shall we weight the centres?, *Statistics in Medicine*, **18**, 365-373.

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

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

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data, *Biometrics*, **54**, 638-645.

Parzen, M., Lipsitz, S. R., & Dear, K. B. G. (1998). Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial?, *Biometrical Journal*, **40**, 385-402.

Patterson, H. D., & Thompson, R. (1971). Maximum likelihood estimation of components of variance. In L. C. A. Corsten, & T. Postelnicu (Eds.), *Proceedings of the 8th international biometric conference*, Editura Academica Republicii Socialite România, Bucureşti, pp. 197-207.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459-496). New York: Marcel Dekker.

SAS Institute Inc. (1996). *Changes and Enhancements through Release 6.1*. Cary, N.C.: Sas Institute Inc.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley & Sons.

Senn, S. (1998). Some controversies in planning and analysing multi-centre trials, *Statistics in Medicine*, **17**, 1753-1765.

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

Sommer, A., Tarwotjo, I., Djunaedi, E., West, K. P., Loeden, A. A., Tilden, R., & Mele, L. (1986). Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial, *Lancet*, **1**, 1169-1173.

SPSS Inc. (1998). *SPSS users guide Base 8.0*, Chicago: SPSS Inc.

Stata Corporation (2001). *Stata User's Guide*. College Station, TX: Stata Press.

Sullivan, L. M., Dukes, K. A. and Losina, E. (1999). An introduction to hierarchical linear modelling, *Statistics in Medicine*, **18**, 855-888.

Ontvangen: 15 juni 2000
Geaccepteerd: 12 maart 2001