ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 341–350

Journal of Clinical Epidemiology

A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies

Mirjam Moerbeek[a,*], Gerard J.P. van Breukelen[b], Martijn P.F. Berger[b]

[a]*Utrecht University, Department of Methodology and Statistics, P.O. Box 80140, 3508 TC Utrecht, The Netherlands*

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

Received 6 November 2001; received in revised form 21 November 2002; accepted 19 December 2002

Abstract

This article reviews three traditional methods for the analysis of multicenter trials with persons nested within clusters, i.e., centers, namely naïve regression (persons as units of analysis), fixed effects regression, and the use of summary measures (clusters as units of analysis), and compares these methods with multilevel regression. The comparison is made for continuous (quantitative) outcomes, and is based on the estimator of the treatment effect and its standard error, because these usually are of main interest in intervention studies. When the results of the experiment have to be valid for some larger population of centers, the centers in the intervention study have to present a random sample from this population and multilevel regression may be used. It is shown that the treatment effect and especially its standard error, are generally incorrectly estimated by the traditional methods, which should, therefore, not in general be used as an alternative to multilevel regression.

Keywords: Nested data; Naive regression; Fixed effects regression; Summary measures; Multilevel regression; Mixed effects regression

1. Introduction

In the health and medical sciences, experiments are conducted to compare different treatments in terms of outcome variables measuring the health or behavior of individuals. In this article we focus on the situation where the data obtained have a nested or hierarchical structure, which means that individuals are nested within clusters. For example, in a clinical trial on the effect of different antipsychotics on the mental health, patients were nested within centers [1]. In a trial where a new approach for the detection and managing of hypertension was studied, patients were nested within family practices [2]. Children were nested within villages in a study on the effect of vitamin A supplementation on childhood mortality in north Sumatra [3], and in a smoking prevention intervention, pupils were nested within classes within schools [4,5]. Outcomes of individuals within the same cluster are likely to be correlated, that is, there will be intracluster correlation.

Data from a smoking prevention intervention [4,5] will be used in this article. To keep things simple we will ignore the nesting of classes within schools leaving two levels of nesting: pupils within classes. Similar but more complicated results hold for three levels of nesting. Of course, the methods presented and conclusions drawn in this article are valid for any kind of experiment where persons are nested within clusters, for instance, multicenter clinical trials with patients nested within clinics. Thus, the reader may replace the words smoking prevention intervention, pupil, and class used in this article by terminology from his/her field of science.

The effect of the smoking prevention intervention on smoking behavior can be estimated and tested with regression, in which the outcome variable is regressed on treatment condition and relevant covariates. In the literature, several types of regression are being used for nested experimental data. Three traditional regression methods are naive regression, fixed effects regression, and regression of summary measures. In the naive regression, pupils are the unit of analysis and their nesting within classes, that is, the dependency among the outcomes of pupils within a class, is ignored. In fixed effects regression, classes are treated as fixed, and their differences are taken into account by dummy coding in the regression equation. Treating classes as fixed implies that statistical inference only takes sampling error at the pupil level into account, not sampling error at the class level, and conclusions are, therefore, limited to the classes in the study. The summary measures method is based upon aggregation of pupil level data within the same treatment condition to the class level, and classes are thus the unit of analysis.

* Corresponding author. Tel.: +31-(0)30-2531450; fax +31-(0)30-2535797.

E-mail address: m.moerbeek@fss.uu.nl (M. Moerbeek).

0895-4356/03/$ – see front matter
doi: 10.1016/S0895-4356(03)00007-6

Multilevel regression [6–9] treats pupils as the unit of analysis, but also takes into account the dependence of outcomes of pupils nested within the same class. The multilevel regression model is also referred to as mixed effects regression, random coefficient model [10], or hierarchical linear model [11], and assumes the classes and pupils to represent random samples from some population of classes and pupils within classes, respectively. Under this assumption class and pupil 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 smoking-prevention interventions should be to produce results not only valid for the classes involved in the experiment, but also for a larger population of classes. In that case, the classes involved in the trial have to represent a random sample from the population of classes, and multilevel analysis is a suitable method of analysis. In practice, there may be good reasons for treating classes as fixed, for instance, when the number of classes in the trial is very small, say less than 10 [9,12]. In this article, however, we will focus on the situation where the classes involved in the trial are treated as a not too small random sample from a much larger population of classes.

Multilevel regression is more complex than the more traditional methods, and consequently, investigators may still want to use these traditional methods, even if they want to generalize the results from their trial to all classes in the population. Therefore, a comparison between the traditional methods and multilevel regression in the context of nested experimental data is relevant. In this article, the relationship between the four methods will be discussed, and it will be shown under which circumstances the traditional methods are acceptable, and when and how they may lead to incorrect results. The comparison made in this article is based on a few regression equations and an illustrative example for (a) the estimator of the treatment effect, and (b) its squared standard error, because these two are of main interest in intervention studies. The comparison is made for continuous outcomes, two levels of nesting, and with randomization at either level. For randomization at the class level, classes will be randomly allocated to the treatment conditions, and all pupils within each class receive the same treatment. For randomization at the pupil level, half of the pupils within each class will be randomly assigned to the treatment group while the others will be allocated to the control group.

Part of the comparison has already been made by others, but has been published fragmentarily in various articles [13–20]. In the present article, these results will be presented systematically, and some gaps in knowledge will be filled up. Again, we want to stress that in this article multilevel regression and more traditional methods for *experimental* data with one posttreatment measurement per person are presented, assuming that the assignment of persons to different conditions is under experimental control. Multilevel regression may also be used for *observation* and/or *longitudinal* studies [21,22].

The remainder of this article is as follows: in Section 2 an example data set of a smoking prevention intervention and two different designs for such trials are given. Naive regression, fixed effect regression, and regression of summary measures are presented in Section 3. Section 4 focuses on multilevel regression. In Section 5, the four methods are used to analyze generated data sets, and it is shown that these methods lead to different results. This difference in results will also be explained using a few simple mathematical expressions in the appendix. In Sections 3 to 5 we assume equal class sizes and no covariates, but in Section 6 these assumptions will be relaxed. In Section 7 some conclusions will be presented.

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 pupil level within each class, and Design 2, where randomization is done at the class level. The latter is often referred to as cluster randomization. For nonvarying class sizes we have a sample of n_2 classes and n_1 individuals per class. In Design 1, $\frac{1}{2}n_1$ pupils per class are randomized to the control group and the others to the treatment group; assume n_1 to be even. In Design 2, $\frac{1}{2}n_2$ classes are allocated to each treatment; assume n_2 to be even, and all pupils within the same class receive the same treatment. A graphical representation of these two designs is given in Figure 1 for four classes. Data on both treatment conditions are available in each class for Design 1, and so the interaction between class and treatment condition can be estimated. This is not possible in Design 2, where data on only one treatment condition are available per class, that is, the data on the other treatment condition are missing by design. So individual level randomization is to be preferred to class level randomization if treatment by class interaction is to be evaluated. Furthermore, randomization at the pupil level results in more efficient estimates of the treatment effect, that is smaller standard error, smaller confidence intervals, and larger power on the test on treatment effect [23]. Randomization at the individual level was done in, for example, a trial on the effect of different antipsychotics on the mental health, with patients nested within centers [1]. In some trials, however, randomization at the individual level is not possible, and Design 2 will be the only alternative. For example, randomization was done at the family practice level rather than at the patient level in a trial where a new approach for the detection and managing of hypertension was studied, because it was recognized that the intervention would not function effectively if some patients in a practice were randomized to the treatment group and others not [2]. In general, certain types of intervention require cluster randomization to prevent treatment group contamination.

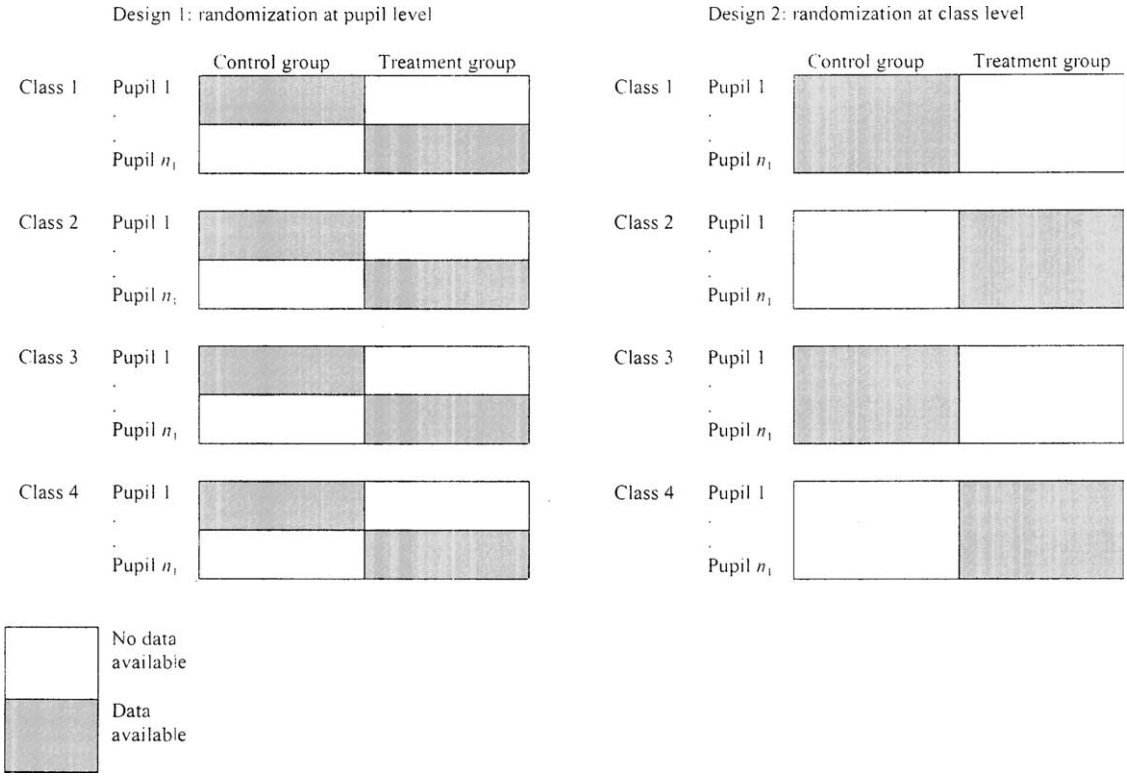

Fig. 1. Graphical representation of Design 1 and Design 2.

The results in this article will be illustrated using a subset of data from the Television School and Family Smoking Prevention and Cessation Project (TVSFP) [4,5]. 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 randomly assigned to one of five treatment conditions. In this article 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 14 schools, 70 classes, and 837 pupils who met these conditions. The dependent variable we used in the analysis is the postintervention Tobacco and Health Knowledge Scale (THKS) score, which was the number of items that a pupil correctly answered in a seven-item questionnaire to assess student tobacco and health knowledge. Although THKS has just eight possible values, we treat it as continuous in our statistical models. This approach is justified as long as the assumption of independence and normality of the error terms are satisfied.

To illustrate the results in Sections 3 to 5 data sets with nonvarying numbers of pupils per class were generated. In Section 6, varying class sizes and the use of covariates will be addressed and the real data will be analyzed.

3. Traditional methods

Three more traditional regression methods for the analysis of multicenter trial data are naïve regression, fixed effects regression, and regression of summary measures. These methods are presented in this section.

3.1. Naive regression

With naïve regression pupils are used as the unit of analysis, and the nesting of pupils within classes is ignored. For both levels of randomization the THKS score denoted y_{ij} and treatment condition denoted x_{ij} of pupil i in class j are related by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + r_{ij}, \quad (1)$$

where the r_{ij} is a random error term assumed to be normally distributed with zero mean and variance, σ_r^2. Naive regression assumes that these error terms are independently distributed, and thus ignored the dependence of THKS scores of pupils within the same class. In this article $x_{ij} = -1$ for the control group and $+1$ for the media group. This is the usual coding scheme for ANOVA and has some advantages compared with (0, 1) coding, namely that β_0 and β_1 are estimated independently, and that simple formulae for the estimators of the regression coefficients and their standard errors are obtained if covariates and/or interactions are added to the

model [24]. So, β_0 is the overall mean of y_{ij} and β_1 is half the difference in outcome between the two treatment conditions. The test statistic for the test of no treatment effect is given by $t = \hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$ with $n_1n_2 - 2$ degrees of freedom assuming independence between all observations. This test is equivalent to an independent samples *t*-test of the mean outcome difference between the treatment and control condition with pupils as unit of analysis.

3.2. Fixed effects regression

In contrast to multilevel regression, fixed effects regression includes class and treatment by class interaction effects as fixed effects, thereby restricting statistical inference to the classes in the study and no other classes. For randomization at the pupil level the n_2 classes may be represented by $n_2 - 1$ "centered" dummy variables, that is, the *j*-th dummy variable equals $+1$ for class *j*, and -1 instead of the usual 0 for the n_2-th class (the so-called "reference class"), and 0 for all other classes. In this way, each dummy variable d_j is on the average 0 within each treatment condition, and each interaction term $x_{ij}d_j$ is uncorrelated with the treatment factor x_{ij} itself. This coding scheme ensures that the regression weights β_0 (intercept) and β_1 of the treatment factor x_{ij} (coded $+1, -1$) retain the same interpretation as in naïve regression, that is, β_0 remains the overall mean outcome, and β_1 half the difference in mean outcome between the two treatments. In fact, due to the balanced design (same n_1 per class, same n_1 and n_2 per treatment) class is not a confounder. Rather, it is included into the model to test for treatment by class interaction and to increase power of the treatment effect test by reducing unexplained variance. The test statistic for testing the treatment effect is calculated as $t = \hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$, and has $n_1n_2 - 2n_2$ or $n_1n_2 - n_2 - 1$ degrees of freedom, for models with or without interaction between classes and treatment, respectively.

For randomization at the class level the same centered coding scheme is used, except that now we need one reference class and $\frac{1}{2}n_2 - 1$ centered dummy variables within each treatment condition (i.e., $n_2 - 2$ dummies overall) to ensure that each dummy is on the average 0 within each treatment so that β_0 and β_1 retain their former interpretation. Of course, due to the nesting of classes within treatment conditions, treatment by class interaction cannot be tested with this design. The test statistic for testing the treatment effect now has $n_1n_2 - n_2$ degrees of freedom.

3.3. Summary measures

The summary measures method tests the treatment effect by using the class means in a *t*-test. So, this method treats the class as unit of analysis, and the outcomes at the pupil level are considered repeated measurements per class that are to be aggregated to class averages. For randomization at the pupil level there are two treatment conditions and two mean outcomes per class, which may be compared using the paired *t*-test. For randomization at the class level there is only one mean outcome per class and the unpaired *t*-test may be used.

4. Multilevel regression

4.1. Design 1: randomization at the pupil level

In multilevel modeling, regression equations are formulated for each level (pupil, class) of the multilevel data structure, and are then combined into a single equation. For randomization at the pupil level, the pupil level equation is given by:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}\,. \tag{2}$$

where e_{ij} is a random error term at the pupil level, and *i* and *j* refer to pupil and class, respectively. Again, the $(-1, +1)$ coding scheme for x_{ij} was used, because of the advantages mentioned in Section 3.1. 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 β_{0j} and slope β_{1j} may vary across classes, randomly and/or as a function of class level covariates. This section will be restricted to equations without any covariate, leaving the inclusion of covariates to Section 6. 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 half the overall treatment effect, respectively. Or, stated differently, u_{0j} is the random class effect and u_{1j} is the random class by treatment interaction effect. Substituting β_{0j} and β_{1j} into equation (2) yields:

$$y_{ij} = \beta_0 + \beta_1x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}\,. \tag{3}$$

The random effects of 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. (The assumption that u_{0j} and u_{1j} are independent of each other stems from so-called mixed and random effects ANOVA, and is not required in multilevel analysis. However, it can be shown that this restriction does not affect estimation and testing of the treatment effect if there is 50:50 randomization to treatment conditions per class and x_{ij} is coded -1 and $+1$ [13], and keeps our presentation more simple.). 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 *j*:

$$\begin{aligned}
&\text{Var}\,(y_{ij}) = \sigma^2_{u0} + \sigma^2_{u1} + \sigma^2_e = \sigma^2_r \\
&\text{Cov}\,(y_{ij}, y_{i'j}) = \sigma^2_{u0} + \sigma^2_{u1} \quad \text{if } x_{ij} = x_{i'j} \\
&\text{Cov}\,(y_{ij}, y_{i'j}) = \sigma^2_{u0} - \sigma^2_{u1} \quad \text{if } x_{ij} \neq x_{i'j}\,,
\end{aligned} \tag{4}$$

with σ^2_r, the error variance as obtained from a naive regression. The so-called intraclass correlation is equal to

$\mathrm{Cov}(y_{ij}, y_{i'j})/\mathrm{Var}(y_{ij})$. If $\sigma^2_{u1} = 0$, this intraclass correlation coefficient reduces to the more familiar expression $\sigma^2_{u0}/(\sigma^2_{u0} + \sigma^2_e)$ which is the proportion of unexplained outcome variance that is at the class level. Because of this intraclass correlation, so-called restricted maximum likelihood or iterative generalized least squares estimation [25–27] instead of ordinary least squares (OLS) regression should be used for parameter estimation, because these estimation methods take the nesting of pupils within classes and the dependence of outcomes of pupils within the same class into account. The null hypothesis of no treatment effect may be tested using the test statistic $t = \hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$ with $n_2 - 1$ degrees of freedom [11]. For a fixed slope β_1 (i.e., without interaction between class and treatment $\sigma^2_{u1} = 0$), the test statistic has $n_1 n_2 - n_2 - 1$ degrees of freedom. It may be noted that this design can also be analyzed with so-called mixed effect factorial ANOVA as implemented in, for example, SPSS, with class as random factor crossed with treatment as fixed factor. For the present design, mixed ANOVA is equivalent to multilevel analysis [24]. However, mixed ANOVA is less suited to designs with varying cluster size and/or covariates.

4.2. Design 2: randomization at the class level

The multilevel regression equation for Design 2 is given by

$$y_{ii} = \beta_0 + \beta_1 x + u_j + e_{ij}\,. \tag{5}$$

where the random interaction term $x_{ij}u_{1j}$ in (3) is omitted because all pupils within the same class receive the same treatment. As a result, σ^2_{u0} associated with the class main effect and σ^2_{u1} associated with the class by treatment interaction cannot be estimated separately. Instead, their sum is estimated, which will be denoted as $\sigma^2_u = \sigma^2_{u0} + \sigma^2_{u1}$ in this article. Furthermore x_{ij} may be replaced by x_j because 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 group. The null hypothesis of no treatment effect is tested by the test statistic $t = \hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$ with $n_2 - 2$ degrees of freedom [11]. This design can also be analyzed with mixed-effect nested ANOVA with classes (random) nested within treatments (fixed), but again, ANOVA is less suited to designs with varying cluster size and/or covariates.

5. Comparison of the four methods

For illustrative purposes we generated a data set with $n_2 = 70$ classes with $n_1 = 12$ pupils each for each level of randomization. We used the parameter values $\beta_0 = 2.34$, $\beta_1 = 0.12$, $\sigma^2_u = 0.16$ and $\sigma^2_e = 1.72$. For randomization at the pupil level the variance σ^2_u was split up into $\sigma^2_{u0} = 0.1$, $\sigma^2_{u1} = 0.06$. These two data sets were analyzed with multilevel regression, naive regression, fixed effects regression, and regression of summary measures. REML estimation as implemented in the computer program MLwiN for multilevel analysis [28] was used for the multilevel analysis, and OLS regression as implemented in SPSS [29] for all other methods. We investigated the assumption of normality of the residuals using normal probability plots and the assumption was found to be satisfied.

The results for Design 1 are given in the upper part of Table 1. We used a model with a random treatment effect since the simulation used a $\sigma^2_{u1} > 0$ as input, implying a random treatment effect (i.e., treatment by class interaction). This table shows that all methods produce the same estimated treatment effect $\hat{\beta}_1$, but that its standard error is underestimated by the naïve regression and fixed effects regression. As a result, the test statistics for these two methods are somewhat larger than those for multilevel analysis and the summary measures method, and *P*-values for fixed effects regression and naïve regression are too small. Note that the summary measures method gives the same $\hat{\beta}_1$, $\hat{SE}(\hat{\beta}_1)$, degrees of freedom and *P*-value (and thus the same results) as multilevel analysis.

The results of the analysis for Design 2 are given in the lower part of Table 1, showing that multilevel analysis and the summary measures method again produce the same estimated treatment effect and standard error, whereas the latter is too small for fixed effects regression and naïve regression.

In the appendix, the results in this section will be explained further using a few simple mathematical expressions. Among others, it shows that, contrary to widespread relief, naïve regression does not always underestimate the standard

Table 1
Results of multilevel and traditional regression analyses of both data sets

	Method			
	Multilevel	Naive	Fixed effects	Summary measures
Design 1: Randomization pupil level				
$\hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$	0.097 (0.050)	0.097 (0.046)	0.097 (0.044)	0.097 (0.050)
$t_{\beta 1}$ (*df*)	1.934 (69)	2.120 (838)	2.205 (700)	1.934 (69)
2-tailed *P*-value	0.056	0.034	0.028	0.056
Design 2: Randomization class level				
$\hat{\beta}_1/\hat{SE}(\hat{\beta}_1)$	0.166 (0.073)	0.166 (0.050)	0.166 (0.048)	0.166 (0.073)
$t_{\beta 1}$ (*df*)	2.285 (68)	3.317 (838)	3.492 (770)	2.284 (68)
2-tailed *P*-value	0.025	0.001	0.001	0.025

error of the treatment effect. For instance, naïve regression leads to an overestimation of this standard error in the situation where randomization is done at the pupil level, and there is no treatment by pupil interaction.

6. Generalization to more complex regression models

The results in the previous section are limited to equal class sizes and regression models with no covariates. Equal class sizes may not be feasible in practice, and often covariates have to be included into the regression model. In this section, these restrictions will be relaxed one at a time. The comparisons are based upon analysis of the TVSP data, with restriction to the Los Angeles pupils in the media or no-treatment control group. Two levels of nesting are taken into account: pupils within classes. Class sizes ranged from 1 to 27, with a mean of 12 pupils per class. Randomization to treatment conditions was done at the school level, so all pupils within a class received the same treatment condition and the interaction between treatment condition and class cannot be estimated.

6.1. Varying class sizes

When the summary measures method is used the classes have to be given a weight in averaging class means. The optimal weights depend on the unknown variance components σ_e^2 and σ_u^2, and are therefore not easily computed. When σ_e^2/n_{1j} is large compared with σ_u^2, the optimal weight for a class is proportional to the number of pupils sampled from that class. On the other hand, if σ_e^2/n_{1j} is small compared with σ_u^2, these weights do not vary across classes (i.e., an unweighted analysis). For fixed effects regression we use dummy variables that are different from those in Section 3.2. Within both treatment conditions the j-th dummy variable equals $+1/n_{1j}$ for class j, $-1/n_{1r}$ for the reference class, and 0 for all other classes, were n_{1j} and n_{1r} are the class sizes in the j-th and reference class, respectively. So, the dummy variables are uncorrelated with treatment and the results from fixed effects regression are equal to those from fixed effects ANOVA.

Both weights are applied in the analysis of the TVSFP data set. Treatment condition was used as the only explanatory variable to model the outcome THKS, leaving the inclusion of the pretreatment THKS as covariate to Section 6.2.

The results are given in the upper half of Table 2. Compared with multilevel analysis, fixed effects regression and naïve regression both produce too small standard errors, and in this specific example also too large $\hat{\beta}_1$. As a result, test statistics are too large and P-values are too small, which was also true for nonvarying class sizes (see Section 5). The estimated treatment effect of fixed effects regression corresponds to that of naïve regression and the summary measures method with weighting according to class size because the dummy variables for fixed class effects are coded such that they are orthogonal to (i.e., uncorrelated with) treatment x_j. Furthermore, the estimated treatment effect and its standard error according to multilevel analysis are bounded by those of the summary measures methods with weighting according to cluster size and without weighting, respectively. In this example an unweighted analysis even produces an estimated treatment effect below zero, due to an extensive variation in class sizes. Normal probability plots of the residual showed that the assumption of normality of the residuals was satisfied.

6.2. Models with covariates

A covariate c_{ij}, for example, age or pretreatment measurement, may be split into a component $\bar{c}_j$, which varies only between classes (e.g., the average pupil age within class j), and a component $c_{ij} - \bar{c}_j$, which varies only within classes [30]. These two components may then be added to the multilevel regression model, the fixed effects regression model, the naïve regression model, and the summary measures method. Note that the class component of the covariate cannot be added to the fixed effects regression model because it is collinear with treatment and the $n_2 - 2$ dummy variables for the fixed class effects. Also, the pupil component of the covariate cannot be added to the summary measures method because it becomes 0 upon aggregation for all classes.

In our TVSFP example the pretreatment THKS was split into a component that varies at the class level and one that varies at the pupil level, and both components were added to the regression models as covariates. As a result, the unexplained variance at both levels decreased. The results of the analyses are given in the lower part of Table 2. Observed P-values were again too low for fixed-effects regression, and naïve regression. Note that the treatment effect estimate according to fixed-effects regression differs from the estimate by the naïve regression, although the dummy variables for class effects are orthogonal to the treatment factor. This is due to the fact that the class level covariate is included in naïve regression but not in fixed effects regression, and this class level covariate correlates with treatment due to sampling error. Again, the estimated treatment effect and its standard error for multilevel regression are bounded by those of the summary measures method with and without weighting. We checked the normal probability plots of the residuals and found that the assumption of normality of the residuals was satisfied.

For all methods except fixed effects regression the estimates of the treatment effect and its standard error differ from those of the models without the covariates $\bar{c}_j$ and $c_{ij} - \bar{c}_j$, because the class level component of the covariate is correlated with treatment condition due to sampling error. Furthermore, including covariates also affects the variance components, and thus also the standard error of the estimated treatment effect. More specifically, $\bar{c}_j$ reduces unexplained variance at the class level and $c_{ij} - \bar{c}_j$ reduces unexplained variance at the pupil level [24]. Although this example is limited to class randomization, similar effects of

Table 2
Results of multilevel and traditional analysis of TVSFP data

	Method				
	Multilevel	Naive	Fixed effects	Summary measures weighting by class size	Summary measures unweighted analysis
Model without pretest THKS					
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.056 (0.070)	0.089 (0.047)	0.089 (0.045)	0.089 (0.067)	−0.041 (0.082)
$t_{\beta 1}$ (*df*)	0.8011 (68)	1.876 (835)	1.964 (767)	1.331 (68)	−0.498 (68)
2-tailed *P*-value	0.426	0.061	0.050	0.188	0.620
Model with pretest THKS					
$\hat{\beta}_1(\hat{SE}(\hat{\beta}_1))$	0.085 (0.061)	0.106 (0.045)	0.089 (0.043)	0.106 (0.060)	−0.018 (0.079)
$t_{\beta 1}$ (*df*)	1.340 (66)	2.369 (833)	2.057 (766)	1.727 (67)	−0.229 (67)
2-tailed *P*-value	0.168	0.018	0.040	0.082	0.819

including covariates can be derived from pupil randomization [24].

7. Conclusions

In this study four methods for the analysis of multilevel experimental data were compared: multilevel analysis, naive regression (persons as unit of analysis), fixed-effects regression, and the use of summary measures (clusters as unit of analysis). 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 in the study, but could also be generalized to the population of clusters. In that case multilevel analysis should be used for the data analysis, but as this method is relatively new and rather complex, it was investigated whether fixed effects regression, naive regression, and the use of summary measures could be used as an alternative to multilevel analysis. As criterion for the comparison the estimator of the treatment effect and its standard error were used, because these are generally of main interest in experimental evaluations of treatments.

The results of the analyses of simulated and real data, and the analytical formulae for the estimator of the treatment effect and its sampling variance (i.e., square standard error) in the Appendix so that naive regression and fixed-effects regression generally result in incorrect estimates of the standard error of the treatment effect, and thereby incorrect results (mostly type I errors) and incorrect confidence intervals (mostly too narrow). Naive regression (i.e., ordinary least squares), which assumes independent outcomes, should only be used if the variance components at the class level are equal to zero, that is, if there are no class effects at all. For varying cluster sizes the use of summary measures without weighting is less efficient than multilevel analysis and weighting by cluster size ignores intraclass correlation. To calculate the correct weights for using summary measures the values of the variance components need to be known. Furthermore, the use of summary measures leads to a loss of information when the model contains covariates. Therefore, multilevel analysis is the only method of the methods discussed in this article that may be used when the study results have to be generalized to some underlying population of clusters from which the clusters in the study are assumed to represent a random sample and cluster sizes vary or covariates are included in the analysis.

There are other advanced methods like Generalized Estimating Equation (GEE [31,32]), and the comparison of this approach with multilevel regression is the subject of several other papers, for example, [22,33]. GEE are often used for longitudinal studies in which repeated measurements on subjects are correlated across time, but may also be used for studies with individuals nested with clusters. GEE is a population-averaged approach, and is used when we want to make inferences about group differences, for instance, differences between treatment groups. Multilevel regression is a subject-specific approach in that it focuses on the change in individuals' responses across time in longitudinal studies. Sometimes the more general term "cluster-specific" is used. An advantage of GEE is that is provides robust estimates of the standard error of the regression coefficients if the correlation structure is miss-specified, which is not necessarily the case for multilevel regression. A disadvantage of GEE is that it cannot easily handle missing data other than those of the Missing Completely at Random (MCAR) type, whereas multilevel regression can also handle the more realistic Missing at Random (MAR) type ([34], p. 218).

This article has systematically compared multilevel regression and traditional approaches, shown which results have also been presented elsewhere, and filled up the gaps in knowledge. Mathematical formulae in the Appendix show that, contrary to the widespread belief, naive regression does not always lead to an underestimation of standard errors of the treatment effect. It should furthermore be noted that, in case of an unbalanced design, multilevel regression not only affects the standard errors of the estimated regression coefficients but also the magnitude of the estimated regression coefficients themselves.

The conventional argument for treatment clusters as fixed or random is whether the results of the study have to be

valid only for the clusters in the sample or for the whole population of clusters, and this argument was used in this article. There are other reasons for using random effects to represent the clusters ([9], pp. 43–44; [27], Section 1.4). For instance, the number of dummy variables in fixed regression is large if the number of clusters involved in the study is large, and one may wish to use multilevel regression instead because with this method only one variance parameter is estimated for the distribution of the clusters. With clusters as fixed effect, the number of dummy indicators and regression parameters increases with the number of clusters sampled. This makes estimation inefficient if not altogether impossible as the number of clusters sampled increases and the number of individuals per cluster decreases.

In this article we focused on analysis methods for multilevel data with continuous outcomes. Comparisons of methods for binary outcomes also show that estimates of regression coefficients, and variance components and their standard errors obtained with multilevel regression may differ from those obtained with traditional methods [35,36]. It should be noted that a number of methods to estimate the regression coefficients and variance components in a multilevel model with dichotomous data are available, but that these methods differ in terms of bias as well as variance of the parameter estimates they provide [37,38].

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. We wish to thank Hubert J. A. Schouten and Martin H. Prins for their comments on this article.

Appendix: comparison of the analysis methods based on mathematical expressions

The four methods will be compared with each other assuming nonvarying class sizes and no covariates. When control and treatment groups are coded by $x_{ij} = -1$ and $x_{ij} = 1$, respectively, the estimator of β_1 is simply half the mean difference in outcome y_{ij} between both conditions for each of the four methods and for both levels of randomization.

In Table 3, the estimated sampling variance (i.e., square standard error) of $\hat{\beta}_1$, $\hat{\text{Var}}(\hat{\beta}_1)$, for the four mthods are given. The second column gives the $\hat{\text{Var}}(\hat{\beta}_1)$ obtained with multilevel analysis. The third column in Table 3 gives the $\hat{\text{Var}}(\hat{\beta}_1)$ obtained with naïve regression, which ignores the nesting of pupils within classes. For randomization at the class level the $\hat{\text{Var}}(\hat{\beta}_1)$ is underestimated with naive regression, and this underestimation depends on the number of pupils per class and the intraclass correlation coefficient $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$, which measures the amount of variance at the class level. For example, if $\rho = 0.1$ and $n_1 = 30$, the confidence interval for β_1 will be about twice as narrow as that obtained with multilevel analysis. For randomization at the pupil level and no treatment by class interaction the $\hat{\text{Var}}(\hat{\beta}_1)$ is slightly overestimated by naïve regression. If treatment by class interaction is present, the $\hat{\text{Var}}(\hat{\beta}_1)$ is under- or overestimated, depending on the values of the variance components and the class size. Thus, naïve regression should not be used for nested data.

Fixed effects regression uses fixed effects (i.e., dummy variables) to represent the classes. The $\hat{\text{Var}}(\hat{\beta}_1)$ obtained with this method is given in the fourth column of Table 3, and does not depend on the level of randomization. For randomization at the class level fixed effects regression underestimates the $\hat{\text{Var}}(\hat{\beta}_1)$, and again, this underestimation depends on the values of ρ and n_1. Even for the small ρ, the underestimation of the $\hat{\text{Var}}(\hat{\beta}_1)$ by fixed-effects regression may be considerable. For example, if $\rho = 0.1$ and $n_1 = 30$, the $\hat{\text{Var}}(\hat{\beta}_1)$ obtained with fixed-effects regression is approximately four times as small as that obtained with multilevel analysis. So the confidence interval for β_1 obtained with fixed-effects regression will be about twice as narrow as that obtained using multilevel analysis, and the null hypothesis of no treatment effect will be rejected too often, leading to an inflation of the type I error rate if the true $\beta_1 = 0$. Thus, fixed

Table 3
Estimated sampling variance (i.e., SE^2) of $\hat{\beta}_1$ for the four regression methods

	Method			
Level of randomization	Multilevel	Naïve	Fixed effects	Summary measures
Pupil (interaction treatment by class)	$\frac{n_1\hat{\sigma}_{u1}^2+\sigma_e^2}{n_1 n_2}$	$\frac{\sigma_r^2}{n_1 n_2} = \frac{\hat{\sigma}_{u0}^2 + \widehat{\hat{\sigma}_{u1}^2} + \sigma_e^2}{n_1 n_2}$	$\frac{\sigma_e^2}{n_1 n_2}$	$\frac{n_1\widehat{\sigma_{u1}^2 + \sigma_e^2}}{n_1 n_2}$
Pupil (no interaction treatment by class)	$\frac{\sigma_e^2}{n_1 n_2}$	$\frac{\sigma_r^2}{n_1 n_2} = \frac{\widehat{\sigma_{u0}^2 + \sigma_e^2}}{n_1 n_2}$	$\frac{\sigma_e^2}{n_1 n_2}$	$\frac{n_1\widehat{\sigma_{u1}^2 + \sigma_e^2}}{n_1 n_2}$
Class	$\frac{n_1\hat{\sigma}_u^2 + \sigma_e^2}{n_1 n_2}$	$\frac{\sigma_r^2}{n_1 n_2} = \frac{\widehat{\sigma_u^2+\sigma_e^2}}{n_1 n_2}$	$\frac{\sigma_e^2}{n_1 n_2}$	$\frac{n_1\widehat{\sigma_u^2 + \sigma_e^2}}{n_1 n_2}$

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 summary measures method the variance components cannot be estimated separately.
Furthermore, for randomization at the pupil level the summary measures method always assumes treatment by class interaction.

Table 4
Comparison of traditional models to the multilevel model with respect to estimated Var($\hat{\beta}_1$) and degrees of freedom for the *t*-test statistic, assuming equal class sizes and classes representing a random sample

	Method		
Level of randomization	Naive (pupils as unit of analysis)	Fixed effects	Summary measures (classes as unit of analysis)
Pupil (interaction treatment by class)	Underestimated or overestimated Var($\hat{\beta}_1$), depending on values variance components	Underestimated Var($\hat{\beta}_1$) [12,14,15].	Correctly estimated Var($\hat{\beta}_1$). Equal to paired samples *t*-test on class by treatment means
	Incorrect df: $n_1n_2 - 2$	incorrect df: $n_1n_2 - 2n_2$	correct df: $n_2 - 1$
Pupil (no interaction treatment by class)	Overestimated Var($\hat{\beta}_1$) [16].	Correctly estimated Var($\hat{\beta}_1$).	Unbiasedly but inefficiently estimated Var($\hat{\beta}_1$)
	Incorrect df: $n_1n_2 - 2$	correct df: $n_1n_2 - n_2 - 1$	unnecessarily low df: $n_2 - 1$
Class	Underestimated Var($\hat{\beta}_1$) [10,19]	Underestimated Var($\hat{\beta}_1$)	Correctly estimated Var($\hat{\beta}_1$). Equal to independent samples *t*-test on class means [18]
	incorrect df: $n_1n_2 - 2$	incorrect df: $n_1n_2 - n_2$	correct df: $n_2 - 2$

effects regression may result in incorrect conclusions, which may also be the case for other values of ρ and n_1 and for pupil level randomization if there is class by treatment interaction. Therefore, it should not be used for the analysis of nested experimental data if the results from the study need to be generalized to some population of classes.

The Vâr($\hat{\beta}_1$) obtained by using summary measures is given in the last column of Table 3. As explained in Section 3.3 this method comes down to performing an independent samples *t*-test (with class randomization) or a paired samples *t*-test (with pupil randomization) with classes as unit of analysis and class averages as observation. For class randomization and for pupil randomization in the presence of interaction, the summary measures method yields the same Vâr($\hat{\beta}_1$) as multilevel analysis. For pupil randomization and no treatment by class interaction it can be shown that the summary measured method is still valid, but less efficient than multilevel analysis [24]. So, for the present examples the use of summary measures is a good alternative to multilevel analysis. However, we do not in general recommend this method as an alternative to multilevel analysis because varying class sizes and the inclusion of covariates into the model are an obstacle to the summary measured method.

The conclusions in this appendix are presented schematically in Table 4. The references show which comparisons have already been made by others, and where their conclusions can be found. The cells without references give conclusions that have not been presented elsewhere, and thus, these gaps in knowledge are filled up in this article.

References

[1] Hedeker D, Gibbons RD, Davis JM. Random regression models for multicenter clinical trials data. Psychopharmacol Bull 1991;27:73–7.

[2] Bass MJ, McWinney IR, Donner A. Do family physicians need medical assistance to detect and manage hypertension? Can Med Assoc J 1986;134:1247–55.

[3] Sommer A, Tarwotjo I, Djunaedi E, West KP, Loeden AA, Tilden R, Mele L. Impact of vitamin A supplementation on childhood mortality. A randomized controlled community trial. Lancet 1986;1:1169–73.

[4] Flay BR, Brannon BR, Johnson CA, Hansen WB, Ulene AL, Whitney-Santiel DA, Gleason LR, Sussman S, Gavin MD, Glowacz KM, Sobol DF, Spiegel DC. The television school and family smoking prevention and cessation project. I. Theoretical basis and program development. Prev Med 1988;17:585–607.

[5] Flay BR, Miller TQ, Hedeker D, Siddiquo O, Britton CF, Brannon BR, Johnson CA, Hansen WB, Sussman S, Dent C. The television, school, and family smoking prevention and cessation project. VIII. Student outcomes and mediating variables. Prev Med 1995;24:29–40.

[6] Goldstein H. Multilevel statistical models. London: Edward Arnold; 1995.

[7] Hox JJ. Multilevel analysis: Techniques and applications. Mahwah, NJ: Erlbaum; 2002.

[8] Kreft I, De Leeuw J. Introducing multilevel modelling. London: Sage Publications; 1998.

[9] Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modelling. London: Sage Publications; 1999.

[10] Longford NT. Random coefficient models. Oxford: Clarendon Press; 1995.

[11] Bryk AS, Raudenbush SW. Hierarchical linear models. Newbury Park, CA: Sage Publications; 1992.

[12] Senn S. Some controversies in planning and analyzing multi-centre trials. Stat Med 1998;17:1753–65.

[13] Raudenbush SW. Hierarchical linear models and experimental design. In: Edward LK, editor. Applied analysis of variance in behavioral science. New York: Marcel Dekker; 1993. p. 459–496.

[14] Gould AL. Multi-centre trial analysis revisited. Stat Med 1998; 17:1779–97.

[15] Jones B, Teather JW, Lewis JA. A comparison of various estimator of treatment difference for a multi-centre clinical trial. Stat Med 1998;17:1767–77.

[16] Parzen M, Lipsitz SR, Dear KGB. Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial? Biometrical J 1998;40:385–402.

[17] Dunlop DD. Regression for longitudinal data: a bridge from least squares regression. Am Stat 1994;48:299–303.

[18] Hopkins KD. The unit of analysis: group means versus individual observations. Am Educ Res J 1982;19:5–18.

[19] Barcikowski RS. Statistical power with group mean as the unit of analysis. J Educ Statistics 1981;6:267–85.

[20] Aitkin M, Longford N. Statistical modelling issues in school effectiveness studies. J R Stat Soc A 1986;149:1–43.
[21] Sullivan LM, Dukes KA, Losina E. An introduction to hierarchical linear modelling. Stat Med 1999;18:855–88.
[22] Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med 1998;17:1261–91.
[23] Moerbeek M, Van Breukelen GJP, Berger MPF. Design issues for multilevel experiments. J Educ Behav Statistics 2000;25:271–84.
[24] Moerbeek M. Design and analysis of multilevel intervention studies. Maastricht: Maastricht University; 2000.
[25] Patterson HD, Thompson R. Maximum likelihood estimation of components of variance. In: Corsten LCA, Postelnicu T, editors. Proc. of the 8th international biometric conference. România, Bucuresti: Editura Academica Republicii Socialite; 1971. p. 197–207.
[26] Goldstein H. Restircted unbiased iterative generalized least squares estimation. Biometrika 1989;76:622–3.
[27] Searle SR, Casella G, McCulloch CE. Variance components. New York: John Wiley & Sons; 1992.
[28] Goldstein H, Rasbash J, Plewis I, Draper D, Brown W, Yang M, Woodhouse G, Healy M. A user's guide to MLwiN. London: Institute of Education; 1998.
[29] SPSS Inc. SPSS user's guide base 8.0. Chicago: SPSS Inc.; 1998.
[30] Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics 1998;54:638–45.
[31] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
[32] Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30.
[33] Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiology 1998;147:694–703.
[34] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.
[35] Hedeker D, McMahon SD, Jason LE, Salina D. Analysis of clustered data in community psychology: with an example from a worksite smoking cessation project. Am J Community Psychol 1994;22: 595–615.
[36] Gibbons RD, Hedeker D. Random effects probit and logistic regression model for three-level data. Biometrics 1997;53:1527–37.
[37] Rodríguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: a case-study. J R Stat Soc A 2001;164:339–55.
[38] Moerbeek M, Van Breukelen GJP, Berger MPF. A comparison of estimation methods for multilevel logistic models. Computat Stat; In press.