Appl. Statist. (2000)
49, *Part* 3, *pp.* 399–412

Meta-analysis using multilevel models with an application to the study of class size effects

Harvey Goldstein and Min Yang

Institute of Education, London, UK

and Rumana Omar, Rebecca Turner and Simon Thompson

Imperial College School of Medicine, London, UK

[Received September 1998. Final revision September 1999]

Summary. Meta-analysis is formulated as a special case of a multilevel (hierarchical data) model in which the highest level is that of the study and the lowest level that of an observation on an individual respondent. Studies can be combined within a single model where the responses occur at different levels of the data hierarchy and efficient estimates are obtained. An example is given from studies of class sizes and achievement in schools, where study data are available at the aggregate level in terms of overall mean values for classes of different sizes, and also at the student level.

Keywords: Class size research; Meta-analysis; Multilevel modelling

1. Introduction

The effects of class size on achievement have been studied since the 1920s quantitatively and qualitatively, and have certainly been debated for much longer. There is a large number of existing studies, including observational surveys, matched designs and randomized controlled trials (RCTs). Despite the number of studies, the results are often inconclusive. Glass and Smith (1979) first applied a meta-analysis to 77 studies based on 70 years' research in more than a dozen countries. They concluded that there were positive effects for class sizes of less than 20, based on 14 of these studies which were considered to be 'well controlled'. Their quantitative synthesis method has been followed by many more meta-analyses on the same topic (Carlberg and Kavale, 1980; Hedges and Olkin, 1985; Slavin, 1986, 1990; McGiverin *et al.*, 1989).

Slavin (1990) argued that Glass's positive finding was based on only a small number of studies and the results were largely affected by one extreme case (Verducci, 1969). On reanalysis Slavin reported an effect that was much smaller than that of Glass. He also conducted an analysis of nine randomized or matched studies. Among these studies some were used by Glass and Smith in 1979 but most of them were new studies selected according to strict inclusion criteria. The large scale Tennessee student/teacher achievement ratio (STAR) RCT study (Word *et al.*, 1990) was included. Slavin suggested a moderate effect size of 0.17 standard deviation (SD) units of achievement score comparing smaller classes of 15 or 16 with larger classes of 25–30.

The use of random-effect models in meta-analysis has been suggested by several researchers (Hedges and Olkin, 1985; Raudenbush and Bryk, 1985; Hardy and Thompson, 1996; Erez *et al.*,

Address for correspondence: Harvey Goldstein, Department of Mathematical Sciences, Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK.
E-mail: hgoldstn@ioe.ac.uk

0035–9254/00/49399

1996; Cleary and Casella, 1997). The present paper focuses more on the methodology of meta-analyses than on the substantive issue of class size *per se*. For a more detailed discussion of the latter and a consideration of the role of RCTs in such studies see Goldstein and Blatchford (1998).

In this paper we tackle the problem of how to compare data from different studies with varying summary measures by using multilevel models (Goldstein, 1995). We also develop multilevel models to combine study level data and individual level data. This provides a statistically efficient method for the situation in which some studies have individual level data but others have only summary statistics available (e.g. means and standard errors from published papers). We first describe, in Section 2, the studies included and data available for addressing the issue of class size effects. Section 3 introduces a multilevel model for meta-analysis, focusing on aggregate level data, and Section 4 describes how the model can be extended to combine both aggregate level and individual level data in the same analysis.

2. Sources of data

2.1. *Criteria for inclusion in the study*

We restrict ourselves to those studies which meet the following inclusion criteria.

(a) The study is an RCT or has a matched design where there is an attempt to match smaller and larger classes initially by using school or student level criteria.
(b) The study outcomes are achievement scores, e.g. standardized test scores or rating scales.
(c) The study is longitudinal with initial and final achievement measures and at least one school year period for both larger and smaller classes.
(d) The smaller class is not less than 15 and the larger class is not more than 40.

These inclusion criteria are similar to those that Slavin (1990) set out for his analysis and the range of class sizes matches that found in educational systems of industrialized countries.

2.2. *Scope and strategy of literature search*

Several databases were searched using the keywords *class size*, *longitudinal study*, *school achievement*; the ERIC database from 1961 to 1997, the *British Education Index* (1954–1996, covering 300 journals of education), the *Canadian Education Index* (1976–1996 coverage) and the *Australian Education Index* (1978–1996 coverage). *Psychological Abstracts* was searched (1985–1996) using the subject titles *class size*, *classroom*, *group size*, *academic achievement* and *meta-analysis*.

Nine studies met our criteria, among which seven studies were used by Slavin (1990). Two studies used by Slavin could not be traced through our database search, or by an additional Internet search for the authors' names. The data on these, as presented by Slavin, are not sufficiently detailed for use in our analysis. Two new studies that were not used by Slavin were added to our collection. Only one study, the STAR study, provides individual level data.

In the next section we list some basic information about the studies selected.

2.3. *Studies selected*

A summary of the statistical information is given in Table 1.

2.3.1. *Study 1: Balow (1969), California*

Study 1 was an experimental (but non-randomized) study on reading achievement for stu-

Table 1. Raw and adjusted data of each study for reading scores

Study *k*	*Grade* *j*	*Class size h*	*Number of pupils,* $n_{h.jk}$	*Mean ± SD reported,* $x_{h.jk} \pm \sigma_{hjk}$	*Adjusted mean* $\hat{\mu}^{C}_{h.jk}$	*Pooled SD,* SD_{jk}	*Standardized adjusted mean* $y_{h.jk}$	*Effect size* $y_{S.jk} - y_{L.jk}$
1	1	15	251	50.9	50.9	12.01†	0.125	
	1	30	744	48.9	48.9	12.01	−0.042	0.17
	3	15	656	248.9	248.9	12.37	0.012	
	3	30	602	245.6	248.6‡	12.37	−0.013	0.02
2	4	16.6	256	0.00 ± 0.30	0.00	0.275	−0.012	
	4	23.7	368	−0.04 ± 0.30	−0.04	0.275	−0.157	0.15
	4	30.3	450	0.02 ± 0.27	0.02	0.275	0.061	−0.07
	4	35.7	555	0.02 ± 0.25	0.02	0.275	0.061	−0.07
3	2	15	78	2.39 ± 0.809	2.67‡	0.620	0.282	
	2	30	542	2.52 ± 0.895	2.47	0.620	−0.041	0.32
	3	15	156	3.16 ± 0.954	3.49‡	0.588	0.212	
	3	30	555	3.42 ± 1.074	3.33	0.588	−0.061	0.27
	4	15	57	4.38 ± 1.181	4.37‡	0.661	0.188	
	4	30	441	4.23 ± 1.400	4.23	0.661	−0.024	0.21
	5	15	43	5.40 ± 1.534	5.55‡	0.665	0.395	
	5	30	413	5.22 ± 1.680	5.21	0.665	−0.041	0.44
	6	15	63	5.69 ± 1.510	6.28‡	0.700	0.320	
	6	30	374	6.19 ± 1.925	6.10	0.700	−0.037	0.36
4	1	15	1127	49.7 ± 14.45	53.4§	9.01	0.098	
	2	25	516	50.6 ± 16.12	50.6	9.01	−0.213	0.31
5	2	15	57	52.0 ± 9.93	52.0	8.379	0.198	
	2	25	55	48.7 ± 7.25	48.7	8.379	−0.205	0.39
6	1	19	368	43.2	43.2	10.46†	−0.085	
	1	31	646	44.6	44.6	10.46	0.049	−0.13
7	1	20	371	523.8 ± 88.7	523.8	137.6	0.191	
	1	27	350	469.8 ± 175.4	469.8	137.6	−0.176	0.39
	2	20	309	590.2 ± 49.6	590.2	50.41	0.115	
	2	27	313	578.7 ± 51.2	578.7	50.41	−0.114	0.23
8	9	20	2819	70.6 ± 11.2§§	70.6	13.3	0.300	
	9	30	2543	62.6 ± 15.4§§	62.6	13.3	−0.301	0.60
9	1	15	2644	531.0 ± 57.1	529.1*	37.23	0.070	
	1	24	1414	520.0 ± 54.4	516.1*	37.23	−0.167	0.24
	2	15	3112	591.0 ± 45.6	591.1*	28.98	0.020	
	2	24	1482	583.0 ± 45.4	579.2*	28.98	−0.042	0.06
	3	15	3353	619.0 ± 38.5	619.5*	21.77	0.008	
	3	24	1357	615.0 ± 38.2	619.2*	21.77	−0.020	0.03

†SD derived from the *F*-test value reported.

‡Both the mean and the SD were adjusted for a pretreatment score based on the reported correlation coefficient between the pretreatment and post-treatment scores.

§Both the mean and the SD were adjusted for a pretreatment score assuming a correlation coefficient of 0.8.

§§Both the mean and the SD were calculated on the basis of an average measure from 10 schools available in the paper.

*Both the mean and the SD were adjusted for a pretreatment score using a three-level model with covariates.

dents from grades 1–2 and then grades 3–4. Class sizes were defined as 15 for small and 30 for large classes. The means of the reading score at grade 1 for the two classes were reported equal so that scores at grade 2 were compared. Means at grade 4 were compared adjusting for the intake reading or pupils' intelligence quotient measured at grade 3. No standard error for any measure was reported, except for *F*-test values in the paper.

2.3.2. *Study 2: Shapson et al. (1980), Toronto city*

Study 2 was an RCT for four class size groups: 16, 23, 30 and 37. The trial period was from grade 4 to grade 5. Efforts were made to keep the same group of pupils in the same class during the trial year. It was reported that the changes in pupils in a class were limited to within ±3 by the end of the study. Measures included test scores for composition, vocabulary, reading, mathematical concepts and mathematical problem solving. Means and SDs adjusted for the year of the study and teachers' experience were reported.

2.3.3. *Study 3: Doss and Holley (1982), Austin, Texas*

Study 3 was a 5-year matched design study from grade 2 to grade 6 for school achievement in reading, language and mathematics. The class size was 15 for small and 30 for large classes. Initial means and SDs of test scores at the beginning of the year and those at the end of the year were reported for the five years. Correlation coefficients between the prescores and post-scores were also reported by grade and class.

2.3.4. *Study 4: Wilsberg and Castiglione (1968), New York City*

A total of 1127 grade 1 students from 13 schools and 516 grade 2 students from seven schools were used in study 4. Grade 1 students were in small classes of 15 and grade 2 students were in large classes of 25 and over. Both received the same materials, and help for a year. The study reported means and SDs of a reading test at entry into the study, and means and SDs of vocabulary and comprehension tests were taken at the end of the study.

2.3.5. *Study 5: Wagner (1981), Toledo, Ohio*

Grade 2 students in one school assigned to small classes of less than 15 were compared with a matched school with large classes of 25 in study 5. This was published as a doctoral thesis.

2.3.6. *Study 6: Mazareas (1981), Boston*

A random sample of 1014 grade 1 pupils (368 from small classes of less than 20 and 646 from large classes of more than 30) were used in study 6. Outcomes were adjusted for covariates and F-test values were reported for five school attainment scores including reading. This was published as a doctoral thesis.

2.3.7. *Study 7: Butler and Handley (1989), Mississippi*

Study 7 was a matched design study of grade 1 and grade 2 students measuring reading, listening and achievement in mathematics. Outcomes for students in smaller classes (size 20) of grade 1 and grade 2 were compared with the same group of students in larger classes (size 27) followed for 2 years. Students in the smaller and larger classes were from the same school. The study matched for factors such as teachers' qualifications and an entrance test, but it did not carry out covariate adjustment. Means and SDs by subject by class group were reported.

2.3.8. *Study 8: San Juan Unified School District (1991), California*

A total of 2819 students from 10 high schools (grade 9) originally in large classes of 30 were assigned to reduced size classes of 20 for a year and compared with those in larger classes in study 8. The means of a reading comprehension test in grades 9 and 10 were reported.

2.3.9. *Study 9: Word et al. (1990), Tennessee*

The STAR project (study 9) was an RCT longitudinal study with children followed from kindergarten to grade 3 for 4 years with measurements at 1-year intervals. Smaller classes averaged about 15 students (13–17) and larger classes about 24 (22–25). About 4000 students were available for the analysis and the initial assignment into kindergarten classes was at random.

In Table 1 the class size, number of pupils, means and SDs are taken from the published papers. The adjusted means and pooled SDs are computed by using equations (1) and (2) respectively below. The standardized adjusted means are computed by using equation (3) below.

As we can see from Table 1 several problems arise. The tests that were used to measure achievement are obviously different from study to study. Rescaling the measurements to a common scale is essential for meta-analysis. Common practice is to standardize the mean for each class group within each study by using a pooled SD. For example, the conventional effect size measure (Glass and Smith, 1979; Hedges and Olkin, 1985) is $(\bar{y}_S - \bar{y}_L)/SD_{\text{pooled}}$, where the terms $\bar{y}_S$ and $\bar{y}_L$ indicate the mean score of smaller and larger class groups respectively. For our purposes we require, as a minimum, estimates of the means and pooled SDs. Some studies did not present SDs for their achievement measures. In this case an F-test or t-test value reported by such a study had to be used to derive the pooled SD for the two groups under comparison.

Differences in the effect of class size between studies may arise from various causes. Where common data are available, e.g. on socioeconomic background, we can see whether such factors explain part of the study differences (Thompson, 1994). In the present case we have the additional problem that different achievement tests were used in each study and this will generally introduce further, unknown, variation. A further issue is that, apart from the STAR study where student level data were available, the between-school variation within a study is not separately reported but should be included in our models.

2.4. Adjusting for pretreatment score

Our inclusion criteria for non-RCT studies to be matched on student or class factors imply that for each study we can adjust for initial achievement. This is important for non-randomized studies to allow for any association between initial achievement and allocation to classes of different sizes. In randomized studies it will generally increase precision as well as potentially helping to correct for any problems with the randomization procedure.

Given the means and SDs for both pretreatment and post-treatment as well as the within-group correlation coefficient r(pre, post) between the pretreatment and post-treatment test scores, we adjust the post-treatment mean of the small and large classes by equation (1) to obtain estimates of the adjusted means $\hat{\mu}^C_{S,\text{post}}$ and $\hat{\mu}^C_{L,\text{post}}$. This is equivalent to applying an analysis of covariance to the two class groups with the pretreatment score as the covariate.

$$\hat{\mu}^C_{h,\text{post}} = \bar{x}_{h,\text{post}} + \frac{\sigma_{\text{post}}}{\sigma_{\text{pre}}} r(\text{pre, post}) (\bar{x}_{h,\text{pre}} - \bar{x}_{\text{pre}}) \tag{1}$$

where h indexes the class size and $\bar{x}_{\text{pre}}$ is the overall pretest mean. The symbols σ and $\bar{x}$ refer to the pooled between-subject SD and treatment means respectively. If the correlation coefficient between the pretreatment and post-treatment scores is not provided, an estimate may be available from other studies (for example see the footnote to Table 1).

2.5. Adjusting and pooling standard deviations

Given the residual sum of squares of the post-treatment score adjusted for the pre-treatment score for each class size group separately, say SS_S and SS_L, a pooled SD is calculated as

$$\mathrm{SD}_{\text{pooled}} = \sqrt{\left(\frac{\mathrm{SS}_\mathrm{S} + \mathrm{SS}_\mathrm{L}}{D_\mathrm{S} + D_\mathrm{L}}\right)} \tag{2}$$

where D refers to the degrees of freedom used for each class size group.

The final summary statistics are the adjusted means and pooled SDs in Table 1. The standardized adjusted means are computed by calculating, for each grade in each study, the mean over all class sizes weighted by the numbers of students, subtracting this from each standardized mean and dividing by the pooled SD, namely

$$y_{h.jk} = \frac{\hat{\mu}^{\mathrm{C}}_{h.jk} - \sum_h n_{h.jk}\hat{\mu}^{\mathrm{C}}_{h.jk} \Big/ \sum_h n_{h.jk}}{\mathrm{SD}_{jk}} \tag{3a}$$

($h = 1$ for a large class; $h = 2$ for a small class). The standardized adjusted means are the responses used for the aggregate level data.

On the basis of these the conventional effect size can be estimated as in the last column of Table 1 by using

$$y_{\mathrm{S}.jk} - y_{\mathrm{L}.jk} = (\hat{\mu}^{\mathrm{C}}_{\mathrm{S}.jk} - \hat{\mu}^{\mathrm{C}}_{\mathrm{L}.jk})/\mathrm{SD}_{jk}. \tag{3b}$$

The homogeneity test (Hedges and Olkin, 1985) for the weighted and bias-corrected effect size estimates for the eight studies with aggregate level data indicates significant heterogeneity between them ($\chi^2_{15} = 255.5$; $p \ll 0.001$). Heterogeneity may have arisen in several ways including inappropriate assumptions about ways of combining effect sizes and omitted levels (between classes and between schools) in the analyses.

As an alternative to working with adjusted effects, we could consider treating the pretest score as a covariate in the multilevel model described in the next section. A difficulty with this approach, however, is that the coefficient for the pretest will vary from study to study, and we shall not pursue this further.

3. A multilevel meta-analysis model

In this section we formulate a general class of meta-analysis models by considering a simple two-level structure. We shall assume that we have a collection of studies, each concerned with the comparison of several 'treatments'. These treatments may be distinct categories (represented by dummy variables) or may be effects represented by regression coefficients or a mixture of the two kinds. The basic models that we shall develop are 'variance component' models but we shall also illustrate a random-coefficient model, and the variance heterogeneity case can also be incorporated (Goldstein (1995), chapter 3).

For the ith subject in the jth study who received the hth treatment, we can write a basic underlying model for outcome y_{hij} as

$$\begin{gathered} y_{hij} = (X\beta)_{ij} + \alpha_h t_{hij} + u_{hj} + e_{hij}, \qquad h = 1, \ldots, H, \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_{hj}, \\ u_{hj} \sim N(0, \sigma^2_{hu}), \qquad e_{hij} \sim N(0, \sigma^2_{he}), \end{gathered} \tag{4}$$

where $(X\beta)_{ij}$ is a linear function of covariates for the ith subject in the jth study, u_{hj} is the random effect of the hth treatment for the jth study and e_{hij} is the random residual of the hth treatment for subject i in study j. The term t_{hij} is a dummy treatment variable (contrasted against a suitable base category) and α_h is the treatment contrast of primary interest. If the treatment dummy variables are replaced by a continuous variable t_{ij} then model (4) becomes

$$y_{ij} = (X\beta)_{ij} + \alpha t_{ij} + u_j + e_{ij}, \qquad j = 1, \ldots, J, \quad i = 1, \ldots, n_j,$$
$$u_j \sim N(0, \sigma_u^2), \qquad e_{ij} \sim N(0, \sigma_e^2).$$

It is also possible to allow the variances within and between studies to be different for each treatment or to vary with the value of a continuous treatment variable, leading to complex variance structures (Goldstein (1995), chapter 3). We can also introduce covariates where data are available and appropriate, and interactions between treatments and covariates. For example, a particular treatment contrast may differ according to the covariate values. We may also relax the normality assumption of the level 1 residuals, e.g. if fitting a generalized linear multilevel model (Goldstein, 1995; Turner *et al.*, 1999).

3.1. Aggregate level data

Consider now the case where model (4) is the underlying model but we only have data by treatment group at the study level. Aggregating to this level we write the mean response as

$$y_{h.j} = (X\beta)_{.j} + \alpha_h t_{h.j} + u_{hj} + e_{h.j} \tag{5}$$

where the dot notation denotes the mean for study *j*. This implies particular constraints, e.g. $\mathrm{var}(e_{h.j}) = \mathrm{var}(e_{hij})/n_{hj}$. A difficulty may arise with the first term in equation (5) since this implies that the mean of the covariate function $(X\beta)_{ij}$ for each study is available.

The corresponding model for the case of a continuous treatment variable is

$$y_{.j} = (X\beta)_{.j} + \alpha t_{.j} + u_j + e_{.j}.$$

3.2. The two-treatment case

Consider the special case of two treatments, $h = 1, 2$. We collapse equation (5) and, using an obvious notation, rewrite to give

$$\begin{aligned} y'_{.j} = y_{1.j} - y_{2.j} &= \alpha + u'_j + e'_{.j}, \\ \alpha &= \alpha_1 - \alpha_2. \end{aligned} \tag{6}$$

This implies the constraint $\mathrm{var}(u'_j) = \mathrm{var}(u_{1j}) + \mathrm{var}(u_{2j}) - 2\,\mathrm{cov}(u_{1j}, u_{2j})$. We can combine equations (5) and (6) into a single model for the case where some aggregated responses are in terms of separate treatment groups and some are in terms of contrasts of groups.

3.3. Defining origin and scale

When combining data from aggregate level studies it is necessary to ensure that the response variable scales are the same and that there is a common origin. In traditional two-treatment meta-analyses the treatment difference is divided by a suitable (pooled) within-treatment SD as described earlier. In our general model, likewise, the response variable in each study can be scaled by dividing it by an estimate of the level 1 SD. Where individual data are available we may use an estimate of the level 1 SD from a preliminary analysis and for aggregate data we may derive this from reported summary information, if this is available.

In situations where the same response variable is used in each study, and scaling has been carried out, we can apply equations (4) and (5) directly. In many cases, however, different response variables are used. For example, in class size studies different reading tests are used. In this case we would not generally expect the means for corresponding treatments to be identical. One procedure for dealing with this is to choose one treatment as a reference treatment (or control) and in each study to subtract its mean from the values of the other treatments and to work *with these differences*. This is the standard approach in two-treatment studies. Thus we chose one treatment described by an intercept term with dummy variables for the remainder. The coefficients of the intercept and of these dummy variables would generally be modelled as random at the study level. In the two-treatment case this leads to model (6).

Where we have a study with individual data we likewise subtracted the mean of the reference treatment group from the response variable. In the fixed part of the model, for the level 1 units with that treatment, the intercept term (and other treatment dummy variables) will be 0.

3.4. Variance information

We may have additional information about variances from studies, e.g. information from other meta-analysis studies about between- or within-study variation. Suppose, for example, that in model (4) we have an external estimate, say r^2_{hue}, of $\sigma^2_{hu} + a^2\sigma^2_{he}$, where we might have $a^2 = 1/n_{hj}$. If we write an additional component to the model as an extra level 2 unit

$$r_{hue} = u_{hj} + ae_{hij} \tag{7}$$

where the fixed part is identically 0 and we have additional constraints imposed as above, this information is then incorporated into the estimation. We note, however, that this extra level 2 unit is given the same weight as every other level 2 unit in the model, and we may wish to assign a different weight depending on the accuracy of the information obtained. Weighting is discussed in the next section.

3.5. Weighting units

We shall consider only weighting of the level 2 units, although extensions to differential weighting of level 1 units are possible. Suppose that the *j*th level 2 unit is assigned a weight w_j. These weights may reflect information about the quality of the study or possibly non-response. Such an analysis might be undertaken as a sensitivity analysis to complement an unweighted analysis. Note that sample size weighting is already incorporated in the estimation via equation (5). Assuming that the weights are uncorrelated with the random effects, we rewrite model (4) to include the vector of the inverses of the square roots of the weights as the explanatory variable for the level 2 random effects. This gives

$$y_{hij} = (X\beta)_{ij} + \alpha_h t_{hij} + u_{hj} w_j^{-0.5} + e_{hij} \tag{8}$$

and we can carry out the standard estimation for this model. This procedure for carrying out a weighted multilevel analysis is discussed in Pfeffermann *et al.* (1997) and is equivalent to their 'step A only' method. They also discussed the case where the weights are correlated with the random effects.

3.6. Modelling class size

In our analysis class size is treated as a continuous variable centred at a value of 15. In all the studies, as is clear from Table 1, only the average class sizes for 'small' or 'large' classes are reported. These values are therefore the values used in the analysis.

One of our aggregate level studies (Doss and Holley, 1982) sampled separate grades within schools. In principle this provides a further level between the class and the school. A preliminary analysis, however, detected variation at this level only for the simplest model, so we do not include it in further models, although grade level itself is incorporated as a fixed factor.

3.7. Aggregate level models for class size data

For the aggregate level studies we can write a basic model as

$$\left.\begin{array}{c} y_{.jk} = \alpha_{0jk} + \alpha_{1k}C_{jk} + \sum_l \beta_l G_{l,jk} + e_{.jk}, \\ \alpha_{0jk} = \alpha_0 + u_{0jk} + v_{0k}, \qquad \alpha_{1k} = \alpha_1 + v_{1k}, \\ u_{0jk} \sim N(0, \sigma^2_{u0}), \qquad e_{.jk} \sim N(0, \sigma^2_e/n_{jk}), \\ v_{0k} \sim N(0, \sigma^2_{v0}), \qquad v_{1k} \sim N(0, \sigma^2_{v1}), \qquad \mathrm{cov}(v_{0k}, v_{1k}) = \sigma_{v01}, \end{array}\right\} \tag{9}$$

where j and k now index the grade and study respectively. The parameter α_0 estimates the mean score for a class size of 15. The term u_{0jk} is the random departure (residual) of the jth-grade mean from the kth study and β_l the fixed effect for grade l, with the $G_{l,jk}$ being grade dummy variables, these being covariates in the model as described in equation (5). The term v_{0k} is the residual for the kth study. The variable C_{jk} is the class size and the parameter α_1 estimates the overall class size effect per additional student. The term v_{1k} estimates the additional random departure for the kth study of the overall class size effect. Further covariates could of course be added, if available. Not all the studies sampled more than one grade level and in some studies several grades are sampled within each school, whereas in others different grades are sampled in different schools. In the latter case grade differences are confounded with school differences so an interpretation of between-grade variation is difficult. For this reason we do not fit grade as a level in the following analysis, although we do study fixed grade effects.

Since all our data have been standardized, the underlying level 1 variance is equal to 1. We therefore define the explanatory variable $z_{jk} = 1/\sqrt{n_{jk}}$ and we can write the first line of model (9) for the aggregated model as

$$\begin{array}{c} y_{.jk} = \alpha_{0jk} + \alpha_{1k}C_{jk} + \sum_l \beta_l G_{l,jk} + w_{jk}z_{jk}, \\ w_{jk} \sim N(0, 1). \end{array} \tag{10}$$

In practice, for classes of a given size in a study, typically we only have available the mean over all classes, so, although the contribution to the variance from these classes for the kth study is $\Sigma_j n_{jk}^{-1}$, the data that are available provide only the value of $(\Sigma_j n_{jk})^{-1}$. When these class sizes are constant, however, the first expression can be obtained from the second where the number of classes is known.

3.8. Results

We first present results for the aggregate level studies only and follow this with results from both the individual level study and the combined individual and aggregate level studies.

Table 2 presents the results of fitting models (9) and (10) for the aggregate data studies (numbers 1–8 in Table 1), using maximum likelihood estimation for three models as shown, together with a 95% confidence interval for the estimates based on a parametric bootstrap with 1000 replications (Goldstein, 1995).

Model A allows the class size effect to vary across studies, model B allows no such variation and model C includes a quadratic effect of class size. As can be seen from the log-likelihoods, model A fits the data substantially better than model B, so there is substantial

Table 2. Model estimates for the aggregated study data using model (9)†

Parameter	*Estimates for the following models:*		
	Model A	*Model B*	*Model C*
Fixed effects			
Intercept	0.163 (0.028, 0.308)	0.207 (0.149, 0.261)	0.224 (0.053, 0.393)
Class size, linear	−0.020 (−0.036, −0.004)	−0.022 (−0.025, −0.019)	−0.048 (−0.072, −0.025)
Class size, quadratic			0.002 (0.001, 0.003)
Random (between-study) effects			
σ^2_{v0}	0.060 (0.0, 0.101)	0.004 (0.0, 0.014)	0.067 (0, 0.135)
σ_{v01}	−0.006 (−0.010, −0.001)		−0.006 (−0.013, 0.004)
σ^2_{v1}	0.0006 (0.0, 0.0010)		0.0006 (0, 0.0010)
−2 log-likelihood	−46.1	266.3	−54.1

†The constrained parameter at class level is omitted. 95% bootstrap intervals are given in parentheses.

evidence of heterogeneity in the class size effect across studies. Model A estimates the effect on reading scores as a decrease of 0.02 SD units per additional student. This is slightly greater than the 0.17 units estimated by Slavin (1990) comparing classes of 15 or 16 with larger classes of 25–30. Model C indicates a quadratic effect of class size whereby from a class size of 15 to a size of 30 there is a continuing decrease in achievement, but an increase in achievement thereafter. This result, however, is influenced by study 2 with the large classes over 30.

A test for equality of grade effects is not significant ($\chi^2_5 = 1.8$) so these have been omitted from these models. The likelihood ratio test statistic suggests that the class size effect varies across studies. However, there are only eight studies in the data set so inferences based on large sample results should be viewed with caution. Also these models ignore between-school variation within studies and between-grade variation as pointed out above. If for model A, however, we allow the level 1 variance to be estimated we obtain an estimate of 1.81 with a likelihood ratio test statistic, for comparison with model A, of 3.0 with 1 degree of freedom so there is only rather weak evidence for a value different from 1.0. If we do the same for model B the level 1 variance estimate is 16.6 and the test statistic is 285.5. The analysis utilizes all the information that is available for the published aggregate studies. Since we are working with standardized data the only flexibility lies in the modelling of the class size effect and the between-study variation. In comparison with the inclusion of individual level data the analysis illustrates the limitations of using aggregate level data.

4. Models for combining individual level data with study level data

Although the STAR individual level data set has covariates available, the aggregate level data have not been adjusted for covariates in a consistent fashion, other than for class size and initial test scores as discussed above. Some studies, however, such as that of Shapson *et al.* (1980), reported their results adjusted for other factors, and some of the studies carried out initial matching. In the following analysis we shall ignore this variation, but it needs to be borne in mind when the results are interpreted.

The STAR study has three levels: school, class and student. Children were recruited when they entered kindergarten where they were randomly assigned to three sizes of class; a small class of 13–17, a regular class of 22–25 and a regular class of 22–25 with a teaching aide. The last two categories are combined since in the STAR study they show no differences. The students were followed for 4 years to the end of grade 3, and for the present purposes we use the reading test score data at the end of grade 1, adjusted for reading test scores at the end of kindergarten, i.e. a study extending over 1 year. The study attempted to retain the original class compositions, but this was not entirely possible. A discussion of the problems of interpreting data from this study is given by Goldstein and Blatchford (1998).

The following model is a combined model for the STAR study and the previously analysed aggregate level studies. We omit the effect of grade since this was not significant for the aggregate level analysis.

$$\left.\begin{array}{c}
y_{ijkl} = (\alpha_{0l} + \alpha_{1l} C_{ijkl}) + e_{.jkl}(1 - z_1) + (\alpha_{0ijkl} + \alpha_2 x_{2ijkl} + v_{1kl} C_{ijkl}) z_l, \\
\alpha_{0l} = \alpha_0 + w_{0l}, \qquad \alpha_{1l} = \alpha_1 + w_{1l}, \qquad \alpha_{0ijkl} = v_{0kl} + u_{0jkl} + e_{ijkl}, \\
z_1 = 1 \text{ if individual data study, } z_1 = 0 \text{ otherwise,} \\
w_{0l} \sim N(0,\, \sigma^2_{w0}), \qquad w_{1l} \sim N(0,\, \sigma^2_{w1}), \qquad \operatorname{cov}(w_{0l},\, w_{1l}) = \sigma_{w01}, \\
v_{0kl} \sim N(0,\, \sigma^2_{v0}), \qquad v_{1kl} \sim N(0,\, \sigma^2_{v1}), \qquad \operatorname{cov}(v_{0kl},\, v_{1kl}) = \sigma_{v01}, \\
u_{0jkl} \sim N(0,\, \sigma^2_{u0}), \qquad e_{ijkl} \sim N(0,\, \sigma^2_e), \qquad e_{.jkl} \sim N(0,\, \sigma^2_e / n_{jkl}),
\end{array}\right\} \quad (11)$$

where x_{2ijkl} is the end of kindergarten score, with the standard assumption that it is independent of the random effects, and C is the class size. The parameter σ^2_{w1} represents the between-study variance in the class size effect, and σ^2_{v1} the between-school variance in the class size effect.

This model utilizes a notation similar to that used before and is now a four-level model with students i grouped within classes j within schools k within studies l. The STAR data are standardized by using the residual variance from a preliminary three-level model with only the STAR data. For the combined data analyses in Table 3 the random-effects parameter estimates at levels 1–3 are derived from the STAR data and at the class level (2) the aggregate level variance, which is not shown, is constrained to be 1. The between-study level (4) intercept and class size coefficient random parameters are estimated from the complete data set.

4.1. Results

In Table 3 the level 4 (between-study) variation is somewhat smaller than that estimated from the aggregate data studies only. We see that the class size effect for the STAR data and the combined estimate is little different from that in the analysis using only aggregate level data (Table 2) and the quadratic effect is now negligible. In fact the linear class size effect in the combined model is less precise than for the STAR study alone because of the substantial heterogeneity between studies in the class size effect. The STAR data show only a small and not significant ($\chi^2_2 = 1.5$) variation in the class size effect between schools. In fact Goldstein and Blatchford (1998) show that for *mathematics* test scores there is a marked variation between schools. A study of the (shrunken) estimated residuals at the study level does not reveal any outliers.

5. Discussion

We have shown how a series of studies, with results reported at different levels of aggregation, can be combined efficiently within a single multilevel model to provide effect size estimates. Since the analysis is based on maximum likelihood estimation within an explicit model it can be expected to yield more efficient estimates than traditional approaches to

Table 3. Parameter estimates for model (11)†

Parameter	*Estimates for the following types of data:*		
	STAR data only	*Combined data (linear)*	*Combined data (quadratic)*
Fixed effects			
α_0	0.078	0.184	0.175
α_1 (class size, linear)	−0.024 (0.006)	−0.022 (0.007)	−0.017 (0.011)
α_3 (class size, quadratic)			−0.0003 (0.0006)
α_2 (pretest)	0.907 (0.018)	0.907 (0.018)	0.907 (0.018)
Random effects			
Level 4 (between study)			
σ^2_{w0}		0.038 (0.020)	0.037 (0.019)
σ_{w01}		−0.004 (0.002)	−0.004 (0.002)
σ^2_{w1}		0.0004 (0.0002)	0.0004 (0.0002)
Level 3 (between school)			
σ^2_{v0}	0.305 (0.064)	0.305 (0.064)	0.305 (0.064)
σ_{v01}	0.00014 (0.004)	0.00012 (0.004)	0.00013 (0.004)
σ^2_{v1}	0.0006 (0.0006)	0.0006 (0.0006)	0.0006 (0.0006)
Level 2 (between class)			
σ^2_{u0}	0.139 (0.023)	0.138 (0.023)	0.138 (0.023)
Level 1 (between student)			
σ^2_e	1.000 (0.023)	1.000 (0.023)	1.000 (0.023)
−2 log-likelihood	11996.5	11948.3	11948.1

†Standard errors are given in parentheses.

meta-analysis. These traditional models also have been unable to combine studies with both individual and aggregate level responses. Our approach does not require balanced data, but it does require that the reporting of studies for inclusion in the model conforms to certain minimum requirements. As we have illustrated, these requirements are such that it should be possible to carry out a suitable standardization for means and variances, after adjusting for relevant covariates. One of the problems with observational studies, especially those involving institutions such as schools, is that (multilevel) modelling incorporating institutional (and other) differences is absent and this can result in biased inferences. In the present case (Table 3) the intraclass and intraschool level correlations are sizable which implies that some of the inferences from the aggregate level studies may overestimate statistical significance. The estimates themselves, however, should be relatively unaffected, and this is consistent with our analysis.

A remaining problem which we have not investigated in detail occurs where studies adjust effects by using different sets of explanatory variables. In the normal distribution case, if information is available about the covariance matrix of all such covariates then for the aggregate level studies common adjustments can be carried out as we have done in model (1).

The model can be extended readily to the multivariate case where more than one outcome is considered, e.g. in the bivariate analysis of mathematics and reading achievement scores. This approach can also be used where not all studies measure all responses so the joint analysis within a single model will provide more efficient estimates than analysing each response separately.

Since we have adopted a model-based approach it is possible in principle to incorporate further model components. An important component is the modelling of publication bias (Copas, 1999), although such models may not lead to improved estimates unless the bias is large (Hedges and Vevea, 1996). In the present case we would argue that publication bias may not be a serious issue. The criteria for study selection have been quite stringent so the relevant studies are carefully executed long-term studies which are unlikely to remain unpublished.

It should be noted that in combining studies for modelling purposes we are making an assumption that the responses used in the various studies are indeed measuring the same thing. In social science applications of meta-analysis this is more problematic than in, say, clinical trials and needs to be borne in mind when interpreting results.

Finally, although the thrust of this paper is methodological, it is of interest that the one large RCT gives an estimate for the class size effect which is very similar to that from the observational studies. This point is pursued further by Goldstein and Blatchford (1998) who also discuss the usefulness of RCTs in this kind of research.

Acknowledgements

The research was funded by the Economic and Social Research Council under the programme for the Analysis of Large and Complex Datasets. We are most grateful to the referees and the Joint Editor for their helpful comments.

References

Balow, I. H. (1969) A longitudinal evaluation of reading achievement in small classes. *Elemen. Engl.*, **46**, 184–187.

Butler, J. M. and Handley, H. M. (1989) Differences in achievement for first and second graders associated with reduction in class size. *18th Mid-south Educational Research Association A. Conf., Little Rock, Nov. 8th–10th.*

Carlberg, C. and Kavale, K. (1980) The efficacy of special versus regular class placement for exceptional children: a meta-analysis. *J. Specl Educ.*, **14**, 295–309.

Cleary, R. and Casella, G. (1997) An application of Gibbs sampling to estimation in meta-analysis: accounting for publication bias. *J. Educ. Behav. Statist.*, **22**, 141–154.

Copas, J. (1999) What works?: selectivity models and meta-analysis. *J. R. Statist. Soc.* A, **162**, 95–109.

Doss, D. and Holley, F. (1982) *A Cause for National Pause: Title I Schoolwide Projects.* Austin: Office of Research and Evaluation.

Erez, A., Bloom, M. C. and Wells, M. T. (1996) Using random rather than fixed effects models in meta-analysis: implications for situational specificity and validity generalisation. *Persnl Psychol.*, **46**, 277–306.

Glass, G. V. and Smith, M. L. (1979) Meta-analysis of research on class size and achievement. *Educ. Evaln Poly Anal.*, **1**, 2–16.

Goldstein, H. (1995) *Multilevel Statistical Models.* London: Arnold.

Goldstein, H. and Blatchford, P. (1998) Class size and educational achievement: a review of methodology with particular reference to study design. *Br. Educ. Res. J.*, **24**, 255–268.

Hardy, R. J. and Thompson, S. G. (1996) A likelihood approach to meta-analysis with random effects. *Statist. Med.*, **15**, 619–629.

Hedges, L. and Olkin, I. (1985) *Statistical Methods for Meta-analysis.* Orlando: Academic Press.

Hedges, L. and Vevea, J. (1996) Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *J. Educ. Behav. Statist.*, **21**, 299–332.

Mazareas, J. (1981) Effects of class size on the achievement of first grade pupils. *Doctoral Dissertation.* Boston University, Boston.

McGiverin, J., Gilman, D. and Tillitski, C. (1989) A meta-analysis of the relation between class size and achievement. *Elem. School J.*, **89**, 47–56.

Pfeffermann, D., Skinner, C. J., Holmes, D., Goldstein, H. and Rasbash, J. (1997) Weighting for unequal selection probabilities in multilevel models. *J. R. Statist. Soc.* B, **60**, 23–40.

Raudenbush, S. and Bryk, A. S. (1985) Empirical Bayes meta-analysis. *J. Educ. Statist.*, **10**, 75–98.

San Juan Unified School District (1991) Class size reduction evaluation: freshman English, spring 1991. *Research Report.* San Juan Unified School District, San Juan.

Shapson, S. M., Wright, E. N., Eason, G. and Fitzgerald, J. (1980) An experimental study of the effects of class size. *Am. Educ. Res. J.*, **17**, 141–152.

Slavin, R. (1986) Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educ. Res.*, **15**, 5–11.

——— (1990) Class size and student achievement: is smaller better? *Contemp. Educ.*, **62**, 6–12.

Thompson, S. G. (1994) Why sources of heterogeneity in meta-analysis should be investigated. *Br. Med. J.*, **309**, 1351–1355.

Turner, R. M., Rumana, R. Z., Yang, M., Goldstein, H. and Thompson, S. G. (1999) Multilevel models for meta analysis of clinical trials with binary outcomes. *Statist. Med.*, to be published.

Verducci, F. (1969) Effects of class size on the learning of a motor skill. *Res. Q.*, **40**, 391–395.

Wagner, E. D. (1981) The effects of reduced class size upon the acquisition of reading skills in grade two. *Doctoral Dissertation.* University of Toledo, Toledo.

Wilsberg, M. and Castiglione, L. V. (1968) *The Reduction of Pupil–Teacher Ratios in Grades 1 and 2 and the Provision of Additional Materials: a Program to Strengthen Early Childhood Education in Poverty Schools, New York, NY.* New York: New York City Board of Education.

Word, E. R., Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N., Folger, J. and Breda, C. (1990) The state of Tennessee's student/teacher achievement ratio (STAR) project. *Technical Report 1985-90*. Nashville, Tennessee State University.