CONFIDENTIAL - Please do not quote without permission

Class size and educational achievement: a methodological review[1]

by

Harvey Goldstein and Peter Blatchford

Institute of Education

University of London

email: h.goldstein@ioe.ac.uk

p.blatchford@ioe.ac.uk

1. Introduction and aims

Possibly more has been written about the effects of class size on performance than on any other single topic in education. Yet despite the number of studies, both experimental and observational, and the number of reviews of such studies, there is still no clear consensus about the extent to which classes of different sizes promote the learning of students. In fact, the class size issue illustrates very clearly many of the important issues in the design and interpretation of quantitative educational research, so that this paper will serve also as a discussion of some general conditions for drawing conclusions from educational research studies. Moreover, many of the issues arise in other areas of research, for example medicine. Because the focus of this paper is methodological we do not attempt a review of existing findings (for which see Blatchford and Mortimore, 1994 and Slavin, 1990), but what we have to say will inevitably affect perceptions of these, especially in the reanalysis of data from the Tennessee Student/teacher Achievement Ratio (STAR) study, a large randomised controlled trial (RCT) which has provided the most important evidence so far.

In the course of this paper we will explore the methodology which has been used to date. We will look at both observational[2] studies and RCTs and through a criticism of existing practice will endeavour to establish criteria for judging the usefulness of different study designs, and by so doing help to inform decisions. In keeping with the bulk of the literature in this area we will concentrate on quantitative research.

In particular we shall carry out a detailed analysis of the STAR project data. Since studies of class size take place within existing educational systems which are organised into complex hierarchical structures with students being grouped within classrooms and the latter grouped within schools, it is appropriate to use multilevel statistical models in the analysis of such data and we will describe the advantages of this approach.

Underlying our discussion is the assumption that the point of class size research is to make statements about causation. By causation we mean the inference that from an ‘effect’ of class

[1] This paper was originally prepared for UNESCO and we are most grateful to that organisation for its encouragement and support.

[2] The term 'observational study' is used to denote research which investigates the characteristics of students, classes etc. *as they exist*, without experimental interventions, and attempts to establish relationships among measurements made on these units.

size estimated by research we can assume that moving children from one class size to another will have a similar effect on achievement. Even with the most carefully controlled study causal interpretations will be difficult, not least because we need to take account of the context in which the research has been carried out; whether the 'effect' may vary across schools, educational systems and other contexts such as social background. For observational studies it is essential to adjust for achievement at the start of the period being studied, and for studies with initial random allocation such adjustment has important advantages in terms of estimation efficiency and interpretation. This is necessary in order to allow for a possibly non-random allocation of students to classes: for example lower achieving children may tend to be allocated to smaller classes if the belief is that smaller classes are advantageous for such children. This requirement for validity rules out from consideration a considerable number of large but purely cross-sectional studies.

In the next sections we look at various aspects of study design and analysis which will form a basis for a critique of existing work. We begin by examining the crucial notion of the target population for a study, that is the schools and classrooms for which some statement about the 'effects' of class size is required. Because of their assumed theoretical methodological advantages we then review the application of randomised controlled trials to studies of class size. This is followed by a review of the methodology of existing studies, including the attempts by several researchers to summarise the results of many hundreds of different studies using 'meta analysis'. We then look at the statistical models which may be appropriate for investigating factors associated with differences among classes of varying sizes, and in the course of this we look at how these models have been applied in practice. This leads into the topic of multilevel models and we show how these can be used to provide greater insights than conventional models. We then briefly discuss cost-benefit analysis aspects of class size research.

Following this we describe our reanalysis of the STAR data for Reading and Mathematics achievement. We show that some of the inferences which have been drawn from these data are unsound and we illustrate how more satisfactory conclusions can be drawn. In some cases we show that different conclusions emerge from the reanalysis, although the overall conclusion of a modest class size effect is not refuted. Finally we summarise our findings and draw conclusions for future research.

2. The measurement of class size

The process of measuring, and indeed defining, class size is problematical. First of all, the actual size of class is not the same as the student-teacher ratio which is measured at the school level by dividing the number of students by the number of full time equivalent teachers. This statistic may provide useful additional information about the resources available for teaching but it is the *experienced* size which is of primary interest. This will vary from day to day and from term to term. The number of students formally on the register of a class may differ from those being taught, for example because of absence. The size of class may vary during the school day as students move between lessons or are withdrawn for particular purposes. At entry into elementary schools there may be particular difficulties, with children entering at different times of year or on a part time basis. There is also the issue, in some areas in some educational systems, of multi-grade classes.

Clearly, therefore, measures of class size taken on just a few occasions during a school year, or those which rely upon the formal size at the start of a school year, may be very poor guides to the actual experiences of students. Ideally, a continuous monitoring of class size is required, which can then be analysed to look for useful summary measures, such as the proportion of time spent in classes of different size. There appears to be little research on this issue, and the

unreliability of those measures which have been used in existing studies may explain some of the failure to observe substantial effects.

3. Target populations

While it may sound obvious, it is often forgotten that any results obtained from a sample apply strictly only to the population of schools and students from which that sample is chosen. If the population sampled is not the target population, then to make any inference to such a population requires additional evidence. In addition, it is usually of interest to study effects on subgroups and also whether there are variations between schools in the sizes of effects. For the purpose of making *causal* inferences this latter issue may be crucial and we shall return to it in a later section. Here we shall raise three important concerns about target populations which seem often to have been ignored in this area of research.

The first issue, which is especially relevant to some of the RCTs, arises from the variation in size and methods of organisation of schools. In the area of elementary or primary schooling, the smallest schools may have classes composed of children in different grades or age groups whereas the largest may have three or more classes for each grade or age group. In the latter case the dynamics of class formation are often complicated in ways which are related to pupil attainment, teacher competence and class size: as we have already suggested, for example, lower attaining children and more experienced teachers may be assigned to smaller classes. Causal inferences will need to take account of this, either by statistical adjustment for prior achievement or by initial randomisation. We shall explore the particular problems associated with randomisation in the next section. In both cases, however, where comparisons are made between classes of different sizes *within the same school*, any conclusions will apply strictly only to large schools. The effect of a given reduction of class sizes within a large school may not be the same as an equivalent change in a small school, especially for particular subgroups such as low attainers. Likewise, a study of small schools where there is just one class for each age group or grade, may detect effects of class size changes which will then strictly apply only to such schools. A further possible complication, which will arise in a RCT, is that the only way to reduce class sizes in small schools is by employing an extra teacher for each class, effectively halving the class size so that more general conclusions about different class size reductions cannot be drawn.

A second issue concerns the inherently historical nature of all social research. Social research tends, indeed is forced, into measuring a real population or subpopulation at one point in time within a particular historical setting. By the time the results are available that context normally will have changed, and some assumptions about the continuity of relationships are necessary. This underlines the necessity to develop theoretically grounded analyses whatever the research is about.

The third issue, one which is endemic throughout social research, is that the institutions or populations which are most accessible for study are often atypical. Thus, for example, because much educational research depends on the co-operation of schools and school boards or authorities, it will often tend to be the better resourced ones which can afford the time to participate in a study. It is difficult to quantify such an effect, but for example in the STAR project (Nye, Achilles et al. 1993) schools were required to agree to participate in the study for four years, had to supply any extra accommodation necessary and undertook not to alter the curriculum during the course of the study. We have little information about how these selection criteria may have excluded particular kinds of schools but it is possible that those excluded may have been more poorly resourced or unable to cooperate for reasons which were associated with the effects which any changes in class size would have had. We shall look in more detail at the role of selection criteria for RCTs in the next section.

4. Randomised controlled trials

Randomisation of subjects to different 'treatments' or experimental situations guarantees that, if the randomised allocation is successful, subsequent comparisons of the treatments for any well defined subgroups can assume that random assignment still obtains. This is important if there are interactions in the data, where differences may vary across subgroups. A problem arises, however, if there are 'compositional' effects. Thus, suppose the 'effect' of class size varies according to the proportion of a particular group in the class, say low attaining children. Then the effect of a reduction in size for classes with high proportions of such children will be different to the effect in classes with low proportions of such children. If randomisation has produced a distribution of this group among classes representative of that in the target population then average conclusions will be justified, even where the compositional variable is not included in any statistical model. This can only be achieved for all possible groups, however, if sampling is strictly with respect to the population of interest. As has already been pointed out this may be very difficult to achieve. Ordinarily we cannot anticipate in advance which factors of this kind may be important, nor can we generally stratify for more than a small number of variables at a time. In such a case randomisation does *not* guarantee that all inferences are applicable to the population of interest.

Those designs where randomisation is within schools face a particular problem. This is because such experiments are 'zero-blind' where the subjects of the experiment, the teachers and even the children know which treatment group they are in and have expectations about the likely effect of the treatment. In medical research such experiments would usually be regarded as difficult to justify because the results may reflect expectations as much as 'real' effects of any treatment. Thus in a study such as STAR the expectations about the effects of class size may be partly responsible for observed effects. In this respect a RCT would seem to have lower validity than a purely observational study. The latter involves no manipulative intervention so that the expectations of participants will not be raised as high so that expectations are less likely to be influential. It is sometimes argued that this 'anticipated expectation' effect should be regarded as a legitimate outcome of a study: even if achievements in small classes are raised simply as a result of teacher expectations then this has practical usefulness. There is, however, a problem with this argument. The effect can only work if practitioners believe that the size of class really matters. Suppose that this is not in fact true, in the sense that practitioners who do not share this belief would not generate an effect. Suppose also that we were able to carry out the research to demonstrate that the effect was merely one which depended upon such a belief. We could then only sustain the anticipated expectation effect by not carrying out the key research study, because once such research had demonstrated the existence of such an effect, it would immediately destroy the belief that a real effect was present and hence the future possibility of anticipated expectation effects occurring. If we wished to rely upon such an effect we could do so only by refusing to carry out the crucial research study or to refrain from publicising its results. To base an educational programme upon such a policy seems somewhat risky, not to say cynical.

A further problem with the within-school design is that there is also a lack of independence across treatments since the teachers and children within a school in different class sizes will interact over time and possibly 'contaminate' the effects of the size differences. As we have pointed out, such effects may be worse in a randomised experiment where awareness of the treatment is heightened compared to an observational study. In one study (Shapson, Wright et al. 1980) over 90% of teachers were found to believe that larger classes produced worse results and this expectation seems to be prevalent in all educational systems.

A design such as STAR, where each school has one or more very small classes and one or more very large classes, may correspond to only a limited number of real populations. Thus,

for example, if *all* class sizes were to be reduced so that all schools had very small classes, the results expected by extrapolating from a study such as STAR might not apply to this new population. In general it is difficult to randomly assign units, whether these be children or classes, so that they function *independently*. The nature of educational systems, and social systems in general, is that the complexity of their structures does not allow us to assume the independent operation of units within them. When an RCT changes such a structure in a research study this implies, in a strict sense, that its conclusions can be accepted, if at all, only for populations with a similar structure. In order to generalise beyond such a structure would require an understanding of the interactions among the units at different levels within a population. In the case of the STAR study this would require an understanding of how the interactions among teachers of different sized classes influenced teaching and learning.

If we have a design where randomisation occurs only at the school level, then this avoids contamination but is then not representative of the real world where, typically, differential sizes do exist within schools, so that the requirement for representativeness is not fulfilled. Of course, we could conceive of a target population where schools have equal class sizes and the results of the study might apply to such a system - but *only* to such a system so that it would again be limited. The one population for which it would be useful is that of single class entry schools.

We see that there are some drawbacks to the use of RCTs in educational research. In particular, educational systems are ‘hierarchical’ structures. Learning takes place in groups: group composition and group dynamics involve interactions among the members of groups which may be important associates of learning. Randomisation, if it eliminates naturally occurring patterns, may tell us something about the effects associated with the groups produced by the randomisation procedure, but this may not be all that is required.

Although we have emphasised the problems of RCTs we do not mean to deny that they may be useful in some situations, although the problem of non-blindness will remain a serious one. A naturally occurring situation where they assume importance is where the existing variation does *not* include the features of interest. Thus, if the educational system being studied has a very uniform distribution of class sizes, intervention would be needed to set up classes of the size we wish to investigate and an RCT would be the appropriate approach. To overcome some of the problems we have described, however, requires a more ‘ruthless’ approach to their use. Thus, to avoid the problems of self selection, once a target population is selected, all eligible units would need to be available for inclusion in the sample: the problem is that schools, unlike cabbages, are actively involved in making autonomous decisions.

5. Questions of causation

Two kinds of questions in this research can be distinguished, the *predictive* and the *descriptive*. The predictive question to which an answer is sought is

- 'If class sizes were reduced by a given amount, what effects would this have on student achievements?'

The descriptive question to which research addresses itself is

- 'Do students in smaller classes happen to have higher achievements?'

Observational studies attempt to address the descriptive question directly, by seeking first to determine what differences exist between achievements in classes of different sizes, and then successively adjusting for factors which may 'explain' observed associations between achievement and class size. In order to sustain a belief in an underlying connection between class size and achievement, by careful data collection and modelling, an observational study

seeks to rule out alternative explanations. It may also look for interactions, that is to establish whether the size of the relationship between class size and achievement varies according the values of other variables. If an enduring relationship can be found then we would want to assume that this establishes 'causality'. In this sense, therefore, the analysis of observational studies can be viewed as an attempt to rule out reasons why an answer to the descriptive question does not also apply to the predictive question.

RCTs *directly* attempt to answer the predictive question by intervening to change class sizes and observing the results. Thus, RCTs also attempt to establish 'causality' but they do this by relying on the random allocation to justify inferences which are correct *on average*. Such average effects may, however, mask interesting and important interactions whereby, for example, the class size effect varies according to initial student achievement or background. In other words, it is important to distinguish between a causal relationship which holds *on average* and a series of *factor specific* causal relationships. Such attempts to contextualise class size effects seem to us to be important. For this reason RCTs should not ignore the potential effects of interactions, and in so doing they will be using the same kinds of procedures, typically the same modelling techniques, as observational studies

6. Existing reviews and methodologies of class size research

There are a number of existing reviews of class size research, for example Glass and Smith (1979), Glass, Cahen et al. (1982), Slavin (1990), Blatchford and Mortimore (1994) and NAHT (1996). The first of these references brings together, via a literature search, some 80 existing studies and then carries out a 'meta analysis' in order to arrive at an overall judgement. The second paper is essentially a critique of the first and also introduces a discussion of the design of the most important of the studies. The third and fourth reviews provide recent comprehensive summaries of the results of class size research, especially on younger children. Before discussing more generally the methodologies employed in class size research studies, we look at the particular methodology of 'meta analysis' and how it has been used to draw conclusions by summarising existing research.

6.1 Meta analyses of class size

The idea behind meta analysis (Hedges and Olkin 1985) is to combine results from a number of studies, all addressing ostensibly the same issue, so as to achieve a consensus result which will also be more precise. For each study a summary 'effect' is estimated. In its simplest form this is just the standardised difference between two groups, say the average achievements of small and large classes. More accurate approaches set up a formal statistical model which allows for the differences between and within studies to vary in more or less complicated fashions.

Consider first a single study where a response or outcome variable, such as an achievement test score, Y, is related to class size, X. We assume that the response has been standardised so that the interpretation of the coefficient β is the same in all studies. A common standardisation is to divide by the between-student standard deviation of the response. A simple model can be written as

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (1)$$

In many studies there will be only two or three different class sizes, nevertheless the interpretation of β as the estimated effect of increasing class size by 1 student still holds. This model is too simple and we should include other variables, most notably a pretest or intake measure, so that the study is longitudinal, as well as student variables to adjust for factors such as socio-economic group. We might also wish to allow a more complex relationship with class

size, for example by including a quadratic term in (1) and possibly interactions with student variables.

When combining the information from several studies the traditional approach is to obtain an estimate of β from each study and then to average these in some suitable manner. A more elaborate procedure would be to model the estimates as a function of the average class size and the range of class sizes in each study and perhaps further variables related to the student, school or teacher characteristics in each study. An example is given by Glass and Smith (1979) who selected 80 studies from a literature search using the simple criterion that they reported data on the relationship between class size and achievement. This approach, however, is not statistically efficient and instead (1) can be extended in the following way by writing

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad (2)$$

where the second subscript *j* identifies the study. The u_j can be interpreted as individual study baseline effects. Equation (2) is now what is known as a *multilevel* model where both the u_j and the e_{ij} are treated as random variables, between studies and between students respectively. The advantage of this formulation is that it is statistically efficient and very flexible. Thus it can be extended to include further variables, measured on students schools or teachers. We can also allow the class size 'effect' to vary across studies, that is we can make it a 'random coefficient' denoted by β_{1j}, and we can incorporate mixtures of studies where in some cases information is available only in terms of an overall class size 'effect' (β_{1j}), and in other cases where there are individual student data available. We shall be using such multilevel models later when we come to look at a reanalysis of the STAR data, and a detailed account can be found in Goldstein (1995).

An important drawback of the approach used by Glass and Smith (1979) is that they treated each study which satisfied their basic criteria for inclusion as having the same weight regardless of the sample size or how many separate comparisons a study contributed. Not only did the studies vary in size, they also varied in quality and, as we have already emphasised, were implemented within different educational systems and contexts. Furthermore, they included only RCTs. Slavin (1990) adopted stringent quality criteria for inclusion in his combined analysis, did not exclude observational studies and finished up with ten studies from which an average 'effect size' was calculated. He concluded that effect sizes were moderate for the achievements studied, and comparable to the later STAR project results.

There appear to be no other studies which have used more careful meta analysis procedures and to date none appear to have made use of multilevel models. Both the Glass and Smith and the Slavin studies used dichotomous weighting functions, either a study was included with a weight of 1 or excluded with a weight of 0. Such a weighting system, however, is inefficient and may also be biased by the use only of extreme weights. An adequate weighting system needs to relate the weights both to the numerical strength of evidence, principally the sample size, and to the quality of the study. Furthermore, each study will have been conducted at a particular historical moment in a particular place. Existing evidence suggests that any effects will be different for different subgroups and in different contexts. This needs to be recognised in the meta analyses by attempting to allow for such study specific factors when modelling the effect sizes. Even where this is difficult to do formally in a statistical model, a qualitative analysis along these lines is important. It is clear that the meta analyses so far conducted are inadequate in these respects: they concentrate on attempting to establish a single overall effect, rather than attempting to assess context specificity and variability.

The conclusion of Slavin (1990), based on comparisons of three RCTs and seven observational studies which matched classes or schools in terms of student characteristics, is that the effect size, for children in the very early years of schooling, is about 0.20 - 0.30 standard deviations and that there is little difference between the RCT results and those from observational studies. This suggests that the procedures used in the observational studies were effective in adjusting for non random assignment.

7. Modelling class size effects

In the previous section we introduced a simple model for studying the effect of class size. In this section we elaborate this model and show how various hypotheses can be studied. We begin by considering a simple design where, within a single school, we have information on some measure of interest from children in classes of different sizes. Following this we extend the model to one involving samples from several schools. We have already discussed inferential problems associated with RCTs and observational studies and we shall return to some of these in our discussion of STAR. In this section we deal only with the specification of the statistical models themselves.

The literature on class size includes a number of suggestions for the ways in which outcomes of interest may be influenced by teacher, school or student factors. For simplicity consider a simple extension of (1), where the model refers to a single school and Z is a student variable[3]

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \qquad (3)$$

where, as indicated earlier, we can interpret β_1 as the 'effect' of class size. This can be elaborated in a number of ways, for example by allowing a nonlinear relationship with class size and an interaction between class size and the student variable, say

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i^2 + \beta_4 x_i z_i + e_i \qquad (4)$$

For example, if Z is a dummy variable for gender, coded 1 for a boy and 0 for a girl then β_4 represents the difference between the class size effects for boys and girls.

In both observational studies and RCTs a class size effect may depend on initial status, for example achievement or attitude measures. In an observational study it is necessary to include such initial measures in order to adjust for the possibility that assignment to different class sizes is related to them, and in a RCT it is important because it will improve statistical efficiency and likewise allow us to investigate whether the effect is related to these measures. Suppose W is an initial achievement measure. We can write

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i^2 + \beta_4 x_i z_i + \beta_5 w_i + \beta_6 g(x_i, w_i) + e_i \qquad (5)$$

where g is a function of the initial achievement and class size. In the simplest case it may just be the product of these two variables but generally it may be more complicated. It represents the possibly differential effect of class size according to initial achievement. A model such as (5) generally will be the simplest realistic model for a useful analysis.

[3] We shall refer to such explanatory variables or 'covariates' in a general way, although in practice they may be continuously distributed, such as a previous test score or discrete, such as gender or ethnic origin group, whence they will be specified in terms of a set of dummy variables. Likewise we shall treat the response variable as continuous and Normally distributed although in practice we may be dealing with, say, binary outcomes such as passing or failing an examination, whence we may wish to specify, say, a logit response model. These differences are technically important but do not affect the general interpretations.

7.1 Incorporating the distribution of initial achievement

Mitchell, Beach et al. (1991) and Preece (1987) suggest that class size effects may be a function of the lowest and highest attainments of students within a class, or equivalently the lowest attainment and the range of attainments. While the range is a measure of the spread of achievement, its value tends to increase as the class size increases, whereas the standard deviation used as a measure of spread does not do so, and may also be a useful predictor. These variables are all measured at the level of the class and can be included directly within (5) together with possible interactions, for example between pretest variability and gender. If achievement within a class is related to the extreme attainments within that class, this may explain some of the existing class size findings since in an RCT there will tend to be more extreme attainments in the larger classes. Mitchell, Beach et al. (1991) explore this possibility, using such functions, on the STAR data and also speculate on how different interactional processes among children and between teachers and children may change with changing class size and how this implies particular mathematical forms of the relationships between achievement and class size. Unfortunately, their analyses are all at the level of the classroom, analysing *average* class scores and gains from one year to the next. The problems associated with such aggregate level analysis are often characterised by the term 'ecological fallacy' (Robinson 1951). The relationships between variables measured on students within classes may be quite different from the corresponding relationships among the classroom *averages* of these variables. Similar difficulties are associated with the interpretations of Preece (1987).

To see this, suppose we write a simple model involving initial attainment as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 \overline{w} + e_i \qquad (6)$$

where $\overline{w}$ is the average of the initial attainment for a class. If we now aggregate up to the class level we obtain a model of the form

$$\overline{y} = \beta_0 + \beta_1 \overline{x} + (\beta_2 + \beta_3)\overline{w} + \overline{e} \qquad (7)$$

and the coefficient of the average initial attainment from the analysis of the class averages is the *sum* of the coefficients of the individual and average effects in (6), whereas we require the *separate* estimates for proper inferences about class size effects. The same problem occurs where there are interactions among student level variables, for example gender and initial attainment or ethnic background, since the class average of the interaction variables is not the same as the interaction of their averages and hence not properly interpretable. Inferences about student level relationships from aggregate level analyses are generally only valid in special cases. Moreover, they are also statistically inefficient, leading to unnecessarily large standard errors, because they do not use the full student level data where available. Nevertheless, the possibility of 'compositional' variables, such as the mean or the variability of prior achievement, being important is an interesting one and in some of our analyses of the STAR data we shall explore this further.

In the next section we shall explore further the importance of multilevel modelling for class size studies where the hierarchical nature of educational systems can be handled. First, however, we look at another suggestion for analysing data where there is no random assignment, nor is there any prior attainment measure which can be used for adjustment.

7.2 Two stage least squares

Akerhielm (1995) suggests that, in the absence of randomisation and information about attainment at the time of allocation to classes, a '2-stage least squares' technique can be used

to obtain unbiased estimates of true class size effects. Essentially the 2-stage least squares model operates as follows.

Suppose that students are allocated to class sizes in a manner which is related to their initial achievement and suppose that we wish to estimate the class size effect as in (3) as the regression coefficient β_1. If x_i is correlated with the residual e_i then the usual regression analysis of (1) will yield a biased estimate of β_1. Since we suppose that x_i is correlated with initial attainment, and since we may suppose that y_i is also strongly correlated with initial attainment, even after adjusting for any relationship with class size, the residual can be expected to be correlated with initial attainment and hence with x_i. This is clearly the case in the 'null' situation when there is no relationship with class size and $\beta_1 = 0$.

One procedure for obtaining an unbiased estimate is to find a so-called instrumental variable, *V*, correlated with *X* but *not* with the residual, and after regressing x_i on v_i then using the *predicted value* of x_i, say $\hat{x}_i = \alpha_0 + \alpha_1 v_i$, in place of x_i to yield an unbiased estimate of β_1 (see for example Johnston, 1972). The main difficulty with this procedure is that it has to satisfy these two requirements, especially since the correlation with *X* should be high if an efficient estimate is required. In the present case the absence of a measure of initial attainment makes the application of this procedure dependent on some strong assumptions which cannot be verified. To illustrate this, suppose initially that class size is unrelated to initial attainment and that the following model holds

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \qquad (8)$$

where z_i is a variable such as socio-economic group. We assume also that this variable is uncorrelated with class size, which might happen if classes have been formed effectively at random. Of course, in practice we would expect many variables to enter into such a model, including school factors, but one will suffice. If we now fit the model as before

$$y_i = \beta_0 + \beta_1 x_i + e_i' \qquad (9)$$

then we do obtain an unbiased estimate of β_1 because of our assumptions which imply that the residual is uncorrelated with the class size. If we now choose an instrumental variable for *X* as before, we require that it be uncorrelated with E', namely with *E and Z*. Since there are many possible variables such as *Z* which may enter into (8) this is clearly a very difficult task. The same issue remains when class size is correlated with initial attainment since we still need to satisfy the requirement that any instrumental variable is uncorrelated with any *Z*.

It is clear therefore, that the use of 2-stage least squares correction procedures does not avoid the need to obtain relevant longitudinal data which allow direct adjustment for any possible allocation based upon the values of such variables associated with the outcome of interest. Nor, of course, does it allow the exploration of 'differential' class size effects whose potential importance we have stressed.

7.3 The choice of response variable

A final general issue concerns the choice of response variable. We are not here concerned with the important substantive issue of how to choose or design a suitable measure, whether cognitive or affective, although this is clearly critical in practice. In the discussion so far we have considered models where the response is measured at the end of 'exposure' to education in a class of a given size with an adjustment for prior attainment both to eliminate non random allocation and to study the relationship of any class size effects with that prior attainment. An alternative approach which has sometimes been advocated is the use of so called 'gain scores',

that is choosing as response the simple difference between a student's prior measurement and his or her final one.

To use this approach both measurements need to be measured on the same, standardised, scale. While this choice of response is possible, and may in some circumstances yield useful interpretations, it will generally still be important to study, for such a response, whether there is any interaction between class size and prior achievement. If, in fact, we include prior attainment in the model then such a model can be made formally equivalent to the model with the final score as response, and the important thing is to include prior attainment. The proposal also suffers from the possible disadvantage of having to standardise both prior and final measurements so that the difference is meaningful, this being unnecessary when the final measurement alone is used as response.

A more complex situation arises when a study extends over several measurement occasions, for example with achievement being measured at the end of each year. In this case we can carry out a series of analyses for each pair of measurements where the earlier one (or several) is used as an adjustment for the later one. An alternative is to standardise the measurements at each occasion and then to treat the data as a series of 'repeated measurements' as in growth curve analysis (Goldstein 1995, Chapter 6), and in addition using the first prior measurement for adjustment. This allows us to study the way in which the mean achievement and, for example the average rate of change of achievement, are related to class size. We shall be using both these forms of analysis when studying the STAR data.

8. Multilevel models

We have already introduced a simple multilevel model in our discussion of meta analysis. Here we set out the use of such a model for the general analysis of class size data, and indeed any data collected on students grouped within schools. We shall also introduce a more general notation to avoid too many symbols and to allow for straightforward generalisations. In this we follow the conventions in Goldstein (1995),

Consider a three level model where an outcome or response measurement is made on students within classes within schools. We assume that, in general, and after adjustment for other variables, there remain average differences among schools and among classes. One of the aims of an analysis might be to explain such differences in terms of other predictor or explanatory variables, but inevitably there will be a residual amount of unexplained variation which needs explicitly to be incorporated into the statistical model. Specifically, consider a measurement of achievement taken at the end of a school year which we relate to class size and to a measure of attainment made at the beginning of the school year. A model for the response for the *i*-th student in the *j*-th classroom in the *k*-th school is

$$y_{ijk} = \beta_0 + \beta_1 x_{1jk} + \beta_2 x_{2ijk} + v_{0k} + u_{0jk} + e_{0ijk} \qquad (10)$$

where x_{1jk} is the class size of the *j*-th class in the *k*-th school and x_{2ijk} is the prior attainment score for the *i*-th student in the *j*-th class in the *k*-th school. The random variables V_0, U_0, E_0 represent the unexplained effects associated with schools, classes and students and are known as residuals. In the standard model they are assumed to have Normal distributions with zero mean and to be mutually independent. Such a model, with three separate random components is an important generalisation of the multiple regression or ordinary least squares model. If there are non-negligible school and class effects then any model which ignores their existence will yield inefficient estimates of coefficients and incorrect significance tests and confidence intervals. Furthermore, the modelling of this residual variation, as we shall see, can be crucial for making causal inferences. For both these reasons, there has been considerable activity

during the last decade devoted to developing new statistical methodology and efficient software for fitting such models.

The model (10) states that the class size effect β_1 is constant across all schools. It is possible, however, that the effect varies from school to school as a result of further, unmeasured, factors. To incorporate this possibility in our model we specify that this coefficient is a random variable at the level of the school giving

$$y_{ijk} = \beta_0 + \beta_{1j} x_{1jk} + \beta_2 x_{2ijk} + v_{0k} + u_{0jk} + e_{0ijk}$$
$$\beta_{1j} = \beta_1 + v_{1k} \qquad (11)$$

where we now have an extra random variable V_1, at the school level. The two school level variables have zero means and each has a variance with a covariance between them. We can write

$$\text{var}(v_{0k}) = \sigma^2_{0k}, \quad \text{var}(v_{1k}) = \sigma^2_{1k}, \quad \text{cov}(v_{0k}, v_{1k}) = \sigma_{01k} \qquad (12)$$

The individual values of the school class size effect residuals v_{1k} can also be estimated having fitted the model and can be used, for example, for identifying schools with extreme values which might be caused by other factors.

Model (11) can be extended in a number of directions. For example it can handle multivariate responses where we have measurements on more than one outcome. It can also handle discrete responses such as pass rates or ordered rating scales and mixtures of continuous and discrete responses.

9. Factors which may explain the effect of class size on educational outcomes

We have examined the link between class size differences on the one hand and educational outcomes on the other. An equally important educational issue involves the identification of factors that might explain any link found. In other words, it is important to ask what factors might *mediate* associations between class size and outcomes. There has been little research that can provide information on this issue. Almost all the studies are from the U.S.A, and doubts exist about the reliability of some of the studies (see Blatchford and Mortimore, 1994). The STAR research was not set up to investigate processes that might *explain* any differences found between small and regular classes. This lack of information is unfortunate because, in its absence, it becomes difficult to offer practical guidance on how to maximise the teaching and learning opportunities provided by having classes of different sizes.

As discussed in Blatchford and Mortimore (1994), knowledge about mediating processes might also help to explain why previous research has not always found a link between class size differences and outcomes. It may be, for example, that when faced with a larger class teachers might alter their style of teaching: they might tend to use more whole class teaching and concentrate more on a narrower range of basic topics. In consequence, children's progress in these areas might not be different (and may even be superior to) children taught in smaller classes. More generally, it may be that when faced with larger classes teachers 'compensate' in a number of ways, for example, by working harder to maximise feedback to individual pupils. If this is true then pupil progress may not be affected adversely, but there may be more covert costs, seen in more teacher stress, lower morale and less opportunities for teacher planning. Another possibility is that some teachers do not alter their teaching to take advantage of smaller classes (as found in Shapson et al, 1980) and it is this that might explain why class size differences have little effect. In order to more closely examine these possibilities, detailed information on classroom processes would be needed.

Although we shall not review the research on mediating factors (for reviews see Blatchford and Mortimore, 1994, Cooper, 1989, NAHT, 1996) we can identify some relevant methodological issues.

First, in the case of both experimental and observational studies one basic objective would be to collect information on classroom processes in order to see if they are affected by class size differences and whether they then affect educational outcomes. To take a simple example, it may be that in larger classes teachers have less opportunity to interact with individual pupils and offer them feedback on their work, and it may be this which explains why children in such classes make less progress. What would be needed here, therefore, would be identification and measurement of the mediating variables - in this case the amount of individual attention and feedback experienced by pupils.

It is important to decide whether a variable is a mediating or an outcome variable and some may play both roles. Pupil's difficult behaviour or difficulties in adjusting to school, for example, may be factors affecting the influence of class size - a teacher in a class with more difficult children may devote less time to the remainder and hence they may make less progress. On the other hand, difficulties of adjustment to school might be chosen as an outcome, in the sense that children's difficulties may be brought into being or exacerbated by larger classes.

Another problem is the difficulty that can be faced in producing reliable and valid measures of mediating processes. In the review by Blatchford and Mortimore (1994) the following factors were identified as likely to be important processes: individualisation of teaching, quality of teaching, curriculum coverage, pupil attention, teacher control and time spent on managing pupils' behaviour, space, pupil morale, and pupil-pupil relations. In some cases measures may be tangible and relatively easily measured - for example, the amount of teacher attention to individual children can be assessed using systematic observation methods, although this is very time consuming (see Blatchford et al, 1987). Other mediating factors may be less easy to use. It is difficult, for example, to measure 'quality' of teaching and adequate measures of teacher morale and stress are difficult to define.

One way of conceiving possible explanatory factors is to divide them, following Mitchell et al (1991), into 'direct' and 'indirect' effects. 'Direct' effects relate to the kind of processes within classrooms that we have been discussing in this section. They include such variables as teaching methods, curriculum coverage, pupil attention, and relationships in class. Mitchell et al also propose a separate set of explanatory factors, which they call 'indirect' explanations. These derive from the spread of pupil abilities within a class and comprise what they call 'class heterogeneity', 'instructional pacing', and student grouping or achievement modelling. There are a number of models that could be drawn on and the reader is referred to Dunkin and Biddle (1974), Bennett (1996), Creemers (1994) and Willms (1992). Assuming that mediating processes can be measured reliably and organised in a conceptual framework then it is possible to incorporate these into the kinds of statistical models we have discussed.

In addition to the difficulties we have already outlined in interpreting results from experimental studies, there may be particular difficulties, for example where teachers are asked to teach in a class of a given size for only a short length of time. In such designs, mediating changes in behaviour and attitudes may be a function of the change itself. This is particularly likely when teachers are studied in artificial situations outside their normal classroom experience.

10. Non cognitive responses

Our discussion has tended to assume that the outcomes of interest are 'cognitive' or 'academic' measures of subject learning in, for example, Mathematics. Since education is about more than

cognitive progression, but is also concerned with values of behaviour, citizenship, tolerance etc., it is relevant to ask whether class size can affect the development of such attributes. Prior to attempting to answer such questions, it is necessary to develop ways of recognising, categorising and generally finding suitable ways of measuring these attributes. There is little in the existing literature, however, which is relevant to such questions, partly because it is generally felt that these things are more difficult to measure and partly because there appears to be relatively little political or public emphasis on studying them. From a *methodological* standpoint it is important to decide whether our discussion about procedures for the study of cognitive measures is equally appropriate for non-cognitive ones.

If agreement can be found about suitable ways of measuring attitudes or behaviour, we see no fundamental distinction between the ways of handling these measures and those we have been discussing. At the simplest level, an attitude may be measured as a binary yes/no attribute which is recognised as being present or absent in a student, or it may be assessed as a grade along a multicategory scale. Such measures can be handled by the same general class of statistical models as we have been describing (Goldstein, 1995). We can introduce baseline or initial attitude measures, as well as other factors such as gender and race. The real difficulty is that of developing suitable measures, and ensuring that they are both reliable and comparable among those who use them.

A major advantage which would accrue from the use of such measures is that they could be used alongside cognitive measures in analyses which studied the interrelationships among them and also the extent to which a change in, say, an attitude measure, affected a cognitive outcome, and vice versa.

11. Cost benefit analysis

The principal focus of this paper is on the methodology for making inferences about the effect of class size. It is, however, worth spending a little time on the economic consequences, because decisions about implementing class size reductions will need to be taken in the knowledge of the relative costs and benefits of competing claims. For example one might save teacher salaries through having fewer teachers with larger classes and use the resources instead on the provision of textbooks. Likewise, if larger classes affect learning partly through a reduction in the physical space available to each student, resources might well be used to increase the space available rather than by reducing the number of students per class. This is a somewhat neglected area of study, partly because there is a scarcity of information about the educational benefits which might accrue from the various alternative measures. It is possible however to set up some simple models and assumptions which might help in understanding the problem.

Jamison (1987) attempts to do this by studying the trade-off between increasing class size by a given amount and the equivalent number, say, of textbooks which could be purchased for the same cost. He illustrates numerically the importance of teacher salaries whereby the lower the salary the greater the increase in class size is required to equate to a given number of textbooks. In other words, in poorly resourced systems where teacher salaries tend to be low, textbooks would seem to be a more effective use of resources where larger classes are associated with poorer achievement and more textbooks are associated with better achievement. He also reports the results of a study of textbook use in a poor country and demonstrates large gains associated with the introduction of such materials.

12. A reanalysis of the STAR data

The STAR study has been referred to several times as providing perhaps the most important evidence about class size during the early years of schooling. Its perceived importance stems from its size, its follow-up of the same children over several years and its randomisation of students and teachers to classes of differing sizes. Children were randomly allocated within each of 79 kindergartens to a 'small' class (13-17 children), or a 'regular' or 'regular with extra teacher aide' class (22-25 students). Unfortunately, the actual class sizes created were not available for analysis. We have already discussed the strengths and limitations of RCTs such as STAR and in the remainder of this paper we shall present a reanalysis of some of the data from that study in order to illustrate the methodological points we have been making and also to see what kinds of inferences can be treated as relatively secure. For our purposes we will look only at Mathematics and Reading achievement throughout the four years of the study. It is, of course, possible that other 'response' variables of interest such as attitudes or self concept ratings will show somewhat different patterns, but this will not alter the general *methodological* conclusions we shall be drawing. Nor, as we have already argued, do we need to adopt substantially different statistical methods.

Our analysis looks at a small number of key explanatory variables. It explores the data through a series of models of increasing complexity in order to illustrate ways in which appropriate statistical modelling can uncover relationships and test causal hypotheses. In the course of the analysis we shall also show how the use of multilevel models to analyse these data are more effective in exploiting their complex structure.

Table 1. Mean Mathematics and Reading score at the end of kindergarten by grade 1 class type. Numbers of children in brackets. Scores are standardised to have zero mean and standard deviation 1.

Mathematics

	Grade 1			
kindergarten	small	regular	missing	Total
small	0.26 (1211)	0.02 (101)	-0.25 (450)	0.12 (2762)
regular	0.00 (231)	0.08 (2705)	-0.35 (1174)	-0.04 (4109)
Total	0.22 (1442)	0.07 (2806)	-0.32 (1624)	0.00 (6871)
Reading				
small	0.25 (1202)	-0.14 (100)	-0.17 (434)	0.12 (1736)
regular	0.00 (227)	0.07 (2668)	-0.33 (1147)	-0.05 (4042)
Total	0.21 (1429)	0.05 (2768)	-0.29 (1581)	0.00 (5778)

12.1 Achievement at the end of kindergarten

At the end of the kindergarten year, some of the students were reallocated to different class sizes. The STAR project (Word, 1990) notes that this was to 'achieve sexual and racial balance and to separate incompatible children'. Table 1 shows the kindergarten class by Grade 1 class for the small and regular class types, with the numbers and mean standardised score at the end of kindergarten. Here and in the subsequent analyses we shall group together the regular and

regular with aide class types since they exhibit few differences. There is an overall difference in favour of the small classes whether classified by kindergarten membership or grade 1 membership. For Mathematics those who were in small kindergarten classes had a lower kindergarten score if they moved to a regular class in grade 1, and likewise for those in regular kindergarten classes who moved to small grade 1 classes. For Reading those who moved from small to regular classes had a larger decrease in kindergarten score than those who moved from regular to small classes. Otherwise, the results are similar for Mathematics and Reading. Those who were lost to the study after kindergarten (24%) had a markedly lower score than those who remained. It seems that a change of class size group after kindergarten tended to happen to those with lower scores and those lost to the study had considerably lower than average scores. We also note that a higher proportion of those in regular kindergarten classes were lost to the study than those in small classes and this may reflect external pressures from parents to remove their children. Such a differential loss may explain some of the subsequent findings about the relative lack of further differences between class sizes following the kindergarten year and underlines the importance of retaining participants in a longitudinal study and also following up those who leave in order to assess their later achievements (see below). It also raises the possibility that, consciously or unconsciously, lower achieving children may have been lost to the experimental small classes as a result of the anticipated benefits which teachers of those classes may have assumed would occur and which then failed to materialise. Also, those teaching regular classes may have tended to reallocate lower achieving children to smaller classes because of their anticipated benefits.

In subsequent analyses we shall present results for separate Mathematics and Reading scores. It is possible to carry out a joint 'bivariate' analysis for both scores which would allow us to investigate the correlation between the scores at the pupil, class and school level, but this is of secondary interest for our present purposes. Also, since almost all children have either both scores or neither, there is little gain in statistical efficiency (smaller standard errors) from a joint analysis.

In later analyses we shall be studying class size differences in grades 1 to 3 for given kindergarten scores. Thus, while there is a differential loss in entering grade 1, if we are prepared to assume that this is random, *given the kindergarten score*, any analyses which also adjust for the kindergarten score will provide valid inferences. Nevertheless, we also find further losses after grade 1. For example, for all those who have a kindergarten score, 20% of those present at grade 1 are not present at grade 2 and these have grade 1 scores which are an average of 0.25 standardised score points lower in Mathematics. The scores have been standardised to have a zero mean and standard deviation of 1.0 separately at each grade using the information for all those present at that grade. The differential loss for those who do have a kindergarten score is reflected in Table 6a (below) in that the average standardised scores for those having a kindergarten score varies from grade to grade. Thus, even carrying out analyses conditional on kindergarten score does not fully eliminate potential biases. If we only analyse the subsequent years' differences unconditionally, that is without adjusting for previous achievement, as in other analyses of the STAR data, then we have the additional biases associated with the dropout after kindergarten.

Table 2 presents a series of multilevel models for the kindergarten score as response and formalises the results of Table 1. There are three levels; the student, the class and the school. The class level variation is that between classes within schools after allowing for the average difference between small and regular classes and the other variables in the fixed part of the model. Likewise for the school and student level variation. The STAR project randomised students at entry to kindergarten, but there was no measurement of initial or 'baseline' achievement. This means that there is no good way to check the success of the randomisation and it also means that comparisons at the end of the kindergarten year are more limited. The

presence of baseline measures would have permitted analyses of differential performance whereby, for example, the class size differences might vary with the baseline achievements themselves which may have potentially important practical implications. Although we do not have such baseline measurements we do know the ages of the students when they started at kindergarten. Since at this age there is a weak positive correlation between achievement and age[4], we have therefore included age on September 1st prior to the start of kindergarten in all our analyses as a partial allowance for intake achievement.

[4] We have no direct estimate of this correlation, but, for example, that between Mathematics test score at the end of Kindergarten and age at start of Kindergarten is 0.11.

Table 2. Kindergarten Mathematics and Reading score by class size in kindergarten and grade 1 and age. Standard errors in brackets

Mathematics				
Fixed	A	B	C	D
Intercept (x_0)	0.00	0.11	0.11	0.12
Age (x_1)	0.29 (0.03)	0.26 (0.03)	0.26 (0.03)	0.26 (0.23)
kindergarten:				
regular - small (x_2)	-0.17 (0.05)	-0.14 (0.05)	-0.14 (0.05)	-0.17 (0.05)
missing - small (x_3)		-0.41 (0.04)	-0.41 (0.04)	-0.39 (0.04)
Grade 1:				
regular - small (x_4)		-0.03 (0.04)	-0.03 (0.04)	-0.18 (0.09)
interaction $x_4 * x_2$				0.18 (0.10)
Random:				
σ^2_{v0}	0.17 (0.03)	0.16 (0.03)	0.15 (0.04)	0.16 (0.03)
σ_{v01}			0.00 (0.03)	
σ^2_{v1}			0.01 (0.03)	
σ^2_{u0}	0.12 (0.01)	0.11 (0.01)	0.11 (0.02)	0.11 (0.01)
σ^2_{e0}	0.70 (0.01)	0.67 (0.01)	0.67 (0.01)	0.67 (0.01)
-2*(log likelihood)	15100.2	14891.9	14891.6	14888.5
Reading				
Fixed				
Intercept (x_0)	0.05	0.16	0.16	0.17
Age (x_1)	0.17 (0.03)	0.15 (0.03)	0.15 (0.03)	0.13 (0.06)
kindergarten:				
regular - small (x_2)	-0.18 (0.04)	-0.12 (0.05)	-0.12 (0.05)	-0.13 (0.06)
missing - small (x_3)		-0.38 (0.04)	-0.38 (0.04)	-0.38 (0.04)
Grade 1:				
regular - small (x_4)		-0.07 (0.04)	-0.07 (0.04)	-0.07 (0.04)
interaction $x_4 * x_2$				0.02 (0.07)
Random:				
σ^2_{v0}	0.17 (0.03)	0.16 (0.03)	0.17 (0.04)	0.16 (0.03)
σ_{v02}			-0.01 (0.03)	
σ^2_{v2}			0.02 (0.03)	
σ^2_{u0}	0.10 (0.01)	0.10 (0.01)	0.10 (0.01)	0.10 (0.01)
σ^2_{e0}	0.72 (0.01)	0.70 (0.01)	0.70 (0.01)	0.70 (0.01)
-2*(log likelihood)	15044.0	14897.2	14896.7	14897.1

In this and subsequent tables the 'base' category is the small class, so that the coefficients estimate the regular-small class difference. In the random part of the model we use the subscript v to denote the school level, u to denote the class level and e to denote the student level. The subscript 0 refers to intercept variation, and 1,2... to various random coefficients. Thus, for example, in this table σ^2_{v0} is the between school intercept variance and σ_{v02} is the covariance at the school level between the intercept and the coefficient of x_2 where it is random at the school level.

In Table 2 we show a number of analyses with different explanatory variables. The age effect is linear and highly significant. From analysis A we see that the difference between the small and regular classes is 0.17 and 0.18 units respectively for Mathematics and Reading, although the relatively large standard errors give a wide 95% confidence interval of 0.08 to 0.26 for Mathematics and 0.11 to 0.27 for Reading. This analysis, although without the age adjustment and the multilevel component, is essentially the analysis carried out by the STAR project team (Word et al., 1990). In analyses B and C we have included a term for those students with data

missing in grade 1 and also a term for the grade 1 class size. For Mathematics there is little suggestion of an effect of the grade 1 class size, but in analysis D, where we have fitted an additional interaction term, this implies that those who change, whether from large to small or vice versa, are those with lower kindergarten scores. The difference between those who are in small classes in both years and those in regular classes in both years is estimated to be 0.17, which is the same difference estimated without fitting the grade 1 class size in analysis A. Likewise, we see that those missing in grade 1 have a markedly lower kindergarten score. For Reading, there is little interaction but a suggestion of an additional reduction for those in regular classes in grade 1 whatever their kindergarten class size. Thus, the central finding of the original STAR analysis for the outcome of kindergarten is broadly confirmed, at least for Mathematics, but it reinforces the need for subsequent analyses to adjust for kindergarten score, as we have already discussed.

The other finding from Table 2 is the absence of any between-school variation in the class size difference. If there had been any substantial variation it would be difficult to draw causal conclusions about the effect of class size since we would have then to explain why such variation occurred. To this extent, therefore, the results are consistent with a causal interpretation, although we should note that there are only 79 schools in the study so that there will be considerable uncertainty attached to any estimate of between-school variability as can be seen in analysis C. We have also looked at the possibility that the between-class variation was different for small and large classes. There was little indication of this for Mathematics but some suggestion in the case of Reading that this variation is smaller for the regular classes. Allowing for this in the model changes the other parameters only very slightly and we do not fit this in subsequent analyses. We also allowed the coefficient of age to vary between schools, and although there was evidence for this ($\chi^2 = 7.4$ for Mathematics and 8.0 for Reading with 2 d.f.) the inclusion of this random coefficient did not change the other estimates appreciably and so we have omitted it from most of the subsequent analyses, although it does reappear in the final set of ‘repeated measures’ analyses.

Table 3. kindergarten Mathematics and Reading score by class size, age, gender, socio-economic status and race in kindergarten. Standard errors in brackets

Mathematics			
Fixed	A	B	C
Intercept (x_0)	-0.22	-0.19	-0.25
Age (x_1)	0.33 (0.03)	0.33 (0.03)	0.33 (0.04)
regular - small class (x_2)	-0.17 (0.05)	-0.20 (0.07)	-0.20 (0.06)
black - white (x_3)	-0.32 (0.04)	-0.34 (0.07)	-0.14 (0.05)
upper - lower SES (x_4)	0.42 (0.03)	0.45 (0.05)	0.44 (0.05)
girls - boys (x_5)	0.14 (0.02)	0.09 (0.04)	0.08 (0.05)
interaction $x_2 * x_3$		0.04 (0.08)	-0.02 (0.06)
interaction $x_2 * x_4$		-0.04 (0.06)	-0.03 (0.06)
interaction $x_2 * x_5$		0.07 (0.05)	0.14 (0.05)
Random:			
σ^2_{v0}	0.16 (0.03)	0.16 (0.03)	
σ^2_{u0}	0.11 (0.01)	0.11 (0.01)	
σ^2_{e0}	0.65 (0.01)	0.65 (0.01)	0.91 (0.02)
-2*(log likelihood)	14649.9	14646.4	16033.6
Reading			
Fixed			
Intercept (x_0)	-0.22	-0.22	-0.26
Age (x_1)	0.21 (0.03)	0.21 (0.03)	0.18 (0.04)
regular - small class (x_2)	-0.18 (0.05)	-0.20 (0.07)	-0.18 (0.06)
black - white (x_3)	-0.20 (0.04)	-0.19 (0.07)	-0.11 (0.05)
upper - lower SES (x_4)	0.47 (0.03)	0.45 (0.05)	0.48 (0.05)
girls - boys (x_5)	0.17 (0.02)	0.15 (0.04)	0.16 (0.05)
interaction $x_2 * x_3$		-0.01 (0.08)	-0.03 (0.06)
interaction $x_2 * x_4$		0.03 (0.06)	0.00 (0.06)
interaction $x_2 * x_5$		0.03 (0.05)	0.06 (0.05)
Random:			
σ^2_{v0}	0.14 (0.03)	0.14 (0.03)	
σ^2_{u0}	0.10 (0.01)	0.10 (0.01)	
σ^2_{e0}	0.67 (0.01)	0.67 (0.01)	0.91 (0.02)
-2*(log likelihood)	14580.4	14579.7	15793.1

In Table 3 we elaborate these basic analyses by including gender, social class and race as covariates. The grade 1 class size has been omitted for simplicity. We see that there is little evidence for interactions between any of these variables and the class size effect. There is, however, a large difference in favour of the white, upper socio-economic group students and girls: the socio-economic and race differences are larger than the class size effect. We have also included the results of fitting a single level model in analysis C to illustrate that failure properly to model higher level variation can result in misleading inferences; in this case that for Mathematics there is an interaction between gender and class size and that there is an underestimation of the effect of race.

So far, therefore, we have clear, albeit relatively small, effects for both Reading and Mathematics of similar sizes, in favour of small classes which are consistent across gender, socio-economic group, race and school. We now go on to study the subsequent progress of the students to see whether this difference changes over time.

12.2 Achievement at the end of grade 1.

Table 4. Grade 1 Mathematics and Reading score by class size, age, gender, socio-economic status, race and kindergarten score. Standard errors in brackets. Scores are standardised to have zero mean and standard deviation 1.

Mathematics			
Fixed	A	B	C
Intercept (x_0)	0.06	0.23	0.19
Age (x_1)	-0.04 (0.04)	-0.01 (0.03)	-0.01 (0.03)
regular - small class (x_2)	-0.29 (0.08)	-0.18 (0.06)	-0.17 (0.06)
black - white (x_3)	-0.34 (0.08)	-0.11 (0.06)	-0.08 (0.06)
upper - lower SES (x_4)	0.38 (0.05)	0.15 (0.04)	0.14 (0.04)
girls - boys (x_5)	-0.01 (0.04)	-0.04 (0.03)	-0.04 (0.03)
kindergarten score		0.58 (0.01)	0.59 (0.02)
kindergarten score s.d.		-0.18 (0.12)	-0.14 (0.12)
interaction x_2 * x_3	-0.11 (0.09)	-0.17 (0.07)	-0.18 (0.07)
interaction x_2 * x_4	0.05 (0.06)	0.08 (0.05)	0.08 (0.05)
interaction x_2 * x_5	0.07 (0.05)	0.03 (0.04)	0.03 (0.04)
Random:			
σ^2_{v0}	0.11 (0.03)	0.09 (0.02)	0.07 (0.02)
σ_{v01}			0.01 (0.01)
σ^2_{v1}			0.02 (0.005)
σ^2_{u0}	0.12 (0.02)	0.12 (0.01)	0.12 (0.01)
σ^2_{e0}	0.62 (0.01)	0.38 (0.01)	0.37 (0.01)
-2*(log likelihood)	10321.1	8389.7	8317.1
Reading			
Fixed			
Intercept (x_0)	-0.27	-0.25 (0.09)	-0.26 (0.08)
Age (x_1)	0.02 (0.07)	0.01 (0.03)	0.01 (0.03)
regular - small class (x_2)	-0.13 (0.07)	-0.04 (0.06)	-0.04 (0.06)
black - white (x_3)	-0.06 (0.07)	0.04 (0.06)	0.07 (0.06)
upper - lower SES (x_4)	0.52 (0.05)	0.28 (0.04)	0.27 (0.04)
girls - boys (x_5)	0.21 (0.04)	0.12 (0.04)	0.13 (0.04)
kindergarten score		0.56 (0.01)	0.59 (0.02)
kindergarten score s.d.		0.09 (0.07)	0.12 (0.07)
interaction x_2 * x_3	-0.21 (0.08)	-0.17 (0.07)	-0.18 (0.07)
interaction x_2 * x_4	-0.06 (0.06)	-0.05 (0.05)	-0.05 (0.05)
interaction x_2 * x_5	0.02 (0.06)	0.01 (0.04)	0.01 (0.04)
Random:			
σ^2_{v0}	0.11 (0.02)	0.10 (0.02)	0.10 (0.02)
σ_{v01}			0.02 (0.01)
σ^2_{v1}			0.02 (0.01)
σ^2_{u0}	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)
σ^2_{e0}	0.66 (0.02)	0.42 (0.01)	0.41 (0.01)
-2*(log likelihood)	10046.7	8345.1	8277.1

The random coefficient at level 3 is for the kindergarten score. The fixed part relationship with kindergarten score is not completely linear, with quadratic and cubic terms being significant at the 1% level. Omitting these, however, does not appreciably alter the remaining estimates and they have been omitted for simplicity. There are no significant interactions between class size and kindergarten score. The mean kindergarten score was also fitted in the fixed part of the model, but for both Maths and Reading the coefficient was small and non significant and the results are not presented.

Table 4 shows analyses with and without adjusting for achievement at the end of the kindergarten year. For Mathematics, if no adjustment is made for the kindergarten score, there is an average class size difference of 0.29 standardised units but no interactions between class size and gender, SES or race. For Reading there is an interaction between class size and race so that for a black child in a regular class there is an estimated decrease of 0.21 score points compared to a white child in a regular class. If the kindergarten score is fitted, however, the class size effect for Mathematics is comparable with that in kindergarten and this shows clearly a continuing effect of class size over and above that in kindergarten. After adjusting for kindergarten score both Mathematics and Reading show an interaction for race and class size, with an estimated decrease of 0.17 score points for black children in regular classes. For Reading, however, for white students there is no additional class size effect once adjustment has been made for kindergarten score. In analysis C we note that there is a between-school variation in the relationship of grade 1 score with kindergarten score which implies that schools differ in terms of the amount of progress students make between kindergarten and grade 1, confirming a common finding from school effectiveness studies. We have included the within-class standard deviation of the kindergarten score as an explanatory variable and there is some suggestion that the greater the variability of kindergarten scores the lower the progress, although not quite statistically significant at the 5% level. The mean kindergarten score was also fitted in the fixed part of the model, but for both Maths and Reading the coefficient was small in all our analyses and non significant and the results are not presented. We also need to remember that this estimate is based upon only those students who were in the STAR study, on average two thirds of those in the classes, so that there will be some measurement errors in the estimate of this standard deviation. This relates to our earlier discussion of study design criteria, namely the importance of having good class level information, sampling whole class units. Finally we note that there is little age at entry effect.

Table 5 looks at the between-school variation in the class size effects.

Table 5. Grade 1 Mathematics and Reading score by class size, age, gender, socio-economic status, race and kindergarten score with class size coefficient random at the school level. Standard errors in brackets

Mathematics		
Fixed	A	B
Intercept (x_0)	0.22	0.22
Age (x_1)	-0.01 (0.03)	-0.01 (0.03)
regular - small class (x_2)	-0.18 (0.06)	-0.19 (0.07)
black - white (x_3)	-0.11 (0.06)	-0.11 (0.06)
upper - lower SES (x_4)	0.15 (0.04)	0.15 (0.04)
girls - boys (x_5)	-0.04 (0.04)	-0.04 (0.04)
kindergarten score	0.58 (0.01)	0.58 (0.01)
kindergarten score s.d.	-0.18 (0.12)	-0.18 (0.12)
interaction $x_2 * x_3$	-0.17 (0.07)	-0.16 (0.07)
interaction $x_2 * x_4$	0.08 (0.05)	0.08 (0.05)
interaction $x_2 * x_5$	0.03 (0.04)	0.03 (0.04)
Random:		
σ^2_{v0}	0.09 (0.02)	0.08 (0.03)
σ_{v01}		0.00 (0.02)
σ^2_{v1}		0.02 (0.03)
σ^2_{u0}	0.12 (0.01)	0.12 (0.02)
σ^2_{e0}	0.38 (0.01)	0.38 (0.01)
-2*(log likelihood)	8389.7	8388.7
Reading		
Fixed		
Intercept (x_0)	-0.25	-0.29
Age (x_1)	0.01 (0.03)	0.01 (0.03)
regular - small class (x_2)	-0.04 (0.06)	-0.04 (0.07)
black - white (x_3)	0.04 (0.06)	0.03 (0.06)
upper - lower SES (x_4)	0.28 (0.04)	0.28 (0.04)
girls - boys (x_5)	0.12 (0.04)	0.12 (0.04)
kindergarten score	0.56 (0.01)	0.56 (0.01)
kindergarten score s.d.	0.09 (0.07)	0.14 (0.07)
interaction $x_2 * x_3$	-0.17 (0.070	-0.15 (0.08)
interaction $x_2 * x_4$	-0.05 (0.050	-0.04 (0.05)
interaction $x_2 * x_5$	0.01 (0.04)	0.02 (0.04)
Random:		
σ^2_{v0}	0.10 (0.02)	0.12 (0.03)
σ_{v01}		-0.04 (0.02)
σ^2_{v1}		0.07 (0.02)
σ^2_{u0}	0.07 (0.01)	0.05 (0.01)
σ^2_{e0}	0.42 (0.01)	0.42 (0.01)
-2*(log likelihood)	8345.1	8335.4

The term σ^2_{v1} is the between-school variance of the class size difference. The mean kindergarten score was also fitted in the fixed part of the model, but for both Maths and Reading the coefficient was small and non significant and the results are not presented.

For Reading there is evidence, with a between-school standard deviation of 0.26, for the class size difference, which implies that for some schools the class size effect is in favour of the large classes and in others in favour of small classes.[5] For Mathematics, however, there is little

[5] If we assume most schools will have a value lying within ± 2 standard deviations this implies a range of from -0.56 to 0.48.

evidence of any substantial between-school variation in the class size difference, although the best estimate yields a standard deviation of 0.15. In both cases there is a substantial standard error associated with the variance estimate so that we need to be careful in drawing conclusions. Nevertheless, the result for Reading illustrates an important methodological issue which can be studied only if multilevel modelling is used: that while there may be a very small average effect (-0.04 in the case of Reading) this does not necessarily imply no effect at all. The existence of differential school effects suggests that there may be key explanatory factors not included in the model. Such factors, possibly related to school organisation or teaching styles, may explain some or all of the class size effect variation, and possibly any average effects also. Thus, as has already been pointed out, the existence of between-school variation means that causal inferences about class size effects are less secure.

12.3 Longitudinal (repeated measures) analysis

The final set of analyses consider the complete set of data from grades 1 - 3, to see if there are further trends over this period, conditional on the kindergarten score. There are further losses and movement in and out of the study over this period with some 2% of students, for example, being present at grade 3 but not grade 2. Of those present at grade 1, 21% have no data on Mathematics or Reading achievement at grade 2 and of those present at grade 2, 14% have none at grade three. Those who 'drop out' after grade 1 and do not return appear to have a particularly low kindergarten score (-0.32 for Mathematics and -0.25 for Reading). The dropouts between grades 1 and 2 from regular classes have somewhat lower kindergarten achievements than those in small classes. (0.08 units for Mathematics and 0.13 for Reading). For the dropouts from grade 2 to grade 3 the reverse is the case with those dropping out of the small classes having lower kindergarten Maths scores (0.12 units for Mathematics and 0.05 for Reading). In the original analyses (Word et al., 1990) only those with complete information were retained in the longitudinal analyses of grades 1 to 3 and in addition there was no adjustment for kindergarten score. The differential dropout implies that such an analysis may be seriously biased and any inferences will need to be treated carefully, in particular those referring to any widening of the gap between small and large classes after kindergarten. As we shall show below, the restriction of including those only with complete data is unnecessary since all students with any data during the period can be retained in the analysis. While this does not eliminate possible biases, it will mitigate them by including all available information, in particular when the analysis adjusts for the kindergarten score as we have already discussed. In some situations it is possible to compensate further for the missingness, but this will not be attempted here.

The model we shall use is often referred to as a repeated measures model. It consists of four levels. Level 4 is that of the school, level 3 the class, level 2 the student and level 1 the repeated measurement occasion within student; in this case grade 1, 2 or 3. The kindergarten score is an explanatory variable (covariate) as before, as is age at entry. We introduce the variable 'year' which takes the values 0, 1, 2 for the three grade years and this becomes an explanatory variable. The intercept term in this model therefore represents the grade 1 year effect and the 'interaction' between the variable 'year' and class size measures the linear trend in the class size difference over grades. We are interested in both the overall class size effect, which is the effect at grade 1 and the trend over time and whether there are further interactions between year and gender, SES and race. The fixed part of the model can be written as follows, where for simplicity we have included just the kindergarten score (X_3), a linear trend term for year (X_1), class size in grade 1 (X_2) and the interaction between class size and year trend (X_4).

$$y_{ijkl} = \beta_0 + \beta_1 x_{1\,jkl} + \beta_2 x_{2\,jkl} + \beta_3 x_{3\,ijkl} + \beta_4 x_{4\,jkl} \qquad (13)$$

Year (X_3) is coded as 0 for grade 1, 1 for grade 2 and 2 for grade 3. Thus the overall class size coefficient (β_2) represents the additional class size effect in grade 1 and the coefficient of the class size interaction with the year trend (β_4) represents the average yearly increase in the class size difference after grade 1. In Tables 5a and 5b we have added a number of further variables including random coefficients.

As already mentioned, in the original STAR analysis (Word et al, 1990) only those students with measurements at all 4 occasions and who remained in the same class size throughout the study, were included in the longitudinal analysis. There are two problems with that approach. Firstly, it is inefficient since the analysis can be carried out using all the available measurements. That is, even where scores are missing at some occasions, the scores at the remaining occasions will provide useful information, and also help to reduce any possible biases. The second problem is that we know from our earlier analyses that those who changed class size after kindergarten tended to have lower kindergarten test scores. This implies not only that we should condition on the kindergarten score but also use the class to which the students belonged *after* kindergarten. We therefore use the class size in grade 1, as in our previous analyses, and ignore the reallocations in subsequent grades. A more detailed analysis would take these reallocations into account also. A discussion of how to make proper allowance for students who change their group or class during a longitudinal study is given by Hill and Goldstein (1997).

We should also mention the problem of scaling test scores in repeated measures models such as these. We have chosen to use 'z' scores, standardised to have the same mean and standard deviation at each occasion. Alternative scalings, for example using age equivalent scores, will generally produce somewhat different results and a sensitivity analysis which explores the results of analyses which use such alternatives will provide useful information.

Table 6a. Repeated measures analysis of grades 1-3 Mathematics score by class size, age, gender, socio-economic status, race and kindergarten Mathematics score. Standard errors in brackets. Scores are standardised to have zero mean and standard deviation 1 separately for those with scores at each grade.

Fixed	A	B	C	D
Intercept (Grade 1) effects:				
Overall	0.15	0.14	0.06	0.15
Age	-0.11 (0.03)	-0.11 (0.03)	-0.11 (0.03)	-0.11 (0.03)
regular - small class (grade 1)	-0.16 (0.04)	-0.24 (0.06)		-0.17 (0.05)
black - white	-0.21 (0.04)	-0.21 (0.04)	-0.21 (0.04)	-0.21 (0.04)
upper - lower SES	0.21 (0.02)	0.21 (0.02)	0.21 (0.02)	0.21 (0.04)
girls - boys	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.02)
kindergarten Maths score	0.62 (0.01)	0.62 (0.01)	0.62 (0.01)	0.62 (0.01)
kindergarten Maths score^2	-0.05 (0.01)	-0.05 (0.01)	-0.05 (0.01)	-0.05 (0.01)
regular - small class (kinder.)		0.09 (0.04)	-0.03 (0.03)	
Trend effects:				
Linear	-0.25 (0.04)	-0.25 (0.04)	-0.22 (0.04)	-0.25 (0.04)
Quadratic	0.03 (0.01)	0.03 (0.01)	0.03 (0.02)	0.03 (0.01)
regular - small class (grade 1)	0.08 (0.03)	0.06 (0.04)		0.08 (0.03)
black - white	0.06 (0.03)	0.06 (0.03)	0.06 (0.03)	0.06 (0.03)
upper - lower SES	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)
girls - boys	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
regular - small class (kinder.)		0.021 (0.028)	0.036 (0.021)	
Random:				
Between schools:				
$\sigma^2_{(4)0}$	0.10 (0.02)	0.10 (0.02)	0.09 (0.02)	0.10 (0.03)
$\sigma_{(4)01}$	-0.02 (0.01)	-0.02 (0.01)	-0.02 (0.01)	-0.02 (0.01)
$\sigma^2_{(4)1}$	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
$\sigma^2_{(4)02}$				-0.01 (0.02)
$\sigma_{(4)12}$				0.00 (0.01)
$\sigma^2_{(4)2}$				0.03 (0.02)
Between classes:				
$\sigma^2_{(3)0}$	0.11 (0.01)	0.11 (0.01)	0.11 (0.01)	0.10 (0.01)
$\sigma_{(3)01}$	-0.05 (0.01)	-0.05 (0.01)	-0.05 (0.01)	-0.05 (0.01)
$\sigma^2_{(3)1}$	0.04 (0.005)	0.04 (0.005)	0.04 (0.005)	0.04 (0.005)
Between children:				
$\sigma^2_{(2)0}$	0.26 (0.01)	0.26 (0.01)	0.26 (0.01)	0.26 (0.01)
Between years:				
$\sigma^2_{(1)0}$	0.16 (0.01)	0.16 (0.01	0.16 (0.01)	0.16 (0.01)
$\sigma_{(1)01}$	0.04 (0.003)	0.04 (0.003)	0.04 (0.003)	0.04 (0.003)
-2*(log likelihood)	20074.7	20067.0	20085.7	20070.6

Year is measured from grade 1 as origin. The fixed part is reported for the intercept (grade 1) and linear trend terms grouped separately. The existence of an average trend in the grade scores results from a differential loss of students when those without kindergarten scores are excluded. The notation $\sigma^2_{(h)l}$, $\sigma_{(h)lm}$ denotes respectively the variance for the *l*-th random coefficient at level *h* and the covariance between the *l*-th and *m*-th random coefficients at level *h*. Thus σ^2_{40} is the variance for the intercept at level 4. Note that there is no between-year variation at level 2 and no quadratic trend in the level 1 variance. Number of students in analysis = 4177; number of Maths measurements over all occasions = 10133; number of students with all four maths measurements = 2665. The subscript 2 at level 4 refers to the class size difference.

We see from analysis A in Table 6a that there is an additional effect of being in a small class in grade 1, but that this tends to be reversed in subsequent grades by virtue of the positive (0.08) linear trend term for class size. In analysis B there is also an additional contribution from the kindergarten class size to the grade 1 effect, reflecting the effect of changes after kindergarten. In analysis C the kindergarten effect on its own is non significant. This latter conclusion is also reached by Word et al (1990) for their longitudinal analyses which use just the kindergarten class type to classify the students. This reinforces our earlier remarks about the importance of taking account of changes which occur over time. The non-white students, the higher SES students and the girls show a gain after grade 1, given their kindergarten scores. For the non-white students this reverses to some extent their poorer performance in grade 1.

In addition, there are changes in class allocation between grades 1 and 3; some 6% of those in small classes in grade 1 changing to regular by grade 3 and 9% vice versa. In addition, those who change tend to have lower grade 1 scores, especially those who move from small to regular classes. When we allow for the grade 3 class size, however, there is no important change to the inferences.

Our final analysis, D, indicates that there is only a very small (non significant) between-school variation in the class size difference. It is of some interest that we have been unable to fit a model where the between-student variance changes over grades, indicating that, given the kindergarten score, students do not make differential progress and thus retain their relative rank positions. We note the large between-occasion variation: the covariance term at level 1 implies that, relatively speaking, the between-occasion variation increases linearly with grade so that there is an increasing variability among students over time. At the teacher level, however, there are differential trends, indicating that the rate of progress varies among teachers. In any further analyses this would be worth following up to see if there are factors which might explain these differences, for example teacher qualifications. It is also possible to estimate teacher level residuals to identify those teachers with very small and very large rates of progress.

Table 6b. Repeated measures analysis of grades 1-3 Reading score by class size, age, gender, socio-economic status, race and kindergarten Reading score. Standard errors in brackets

Fixed	A	B	C	D
Intercept (Grade 1) effects:				
Overall	-0.02	-0.02	-0.06	-0.01
Age	-0.14 (0.03)	-0.14 (0.03)	-0.14 (0.03)	-0.15 (0.03)
regular - small class (grade 1)	-0.10 (0.04)	-0.13 (0.05)		-0.10 (0.04)
black - white	-0.04 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.04 (0.04)
upper - lower SES	0.24 (0.03)	0.24 (0.03)	0.24 (0.03)	0.24 (0.03)
girls - boys	0.12 (0.02	0.12 (0.02)	0.12 (0.02)	0.12 (0.02)
kindergarten Maths score	0.69 90.01)	0.69 (0.01)	0.69 (0.01)	0.69 (0.01)
kindergarten Maths score^2	-0.07 (0.01)	-0.07 (0.01)	-0.07 (0.01)	-0.07 (0.01)
regular - small class (kinder.)		0.05 (0.04)	-0.03 (0.03)	
Trend effects:				
Linear	-0.13 (0.03)	-0.13 (0.03)	-0.11 (0.0)	-0.13 90.03)
Quadratic	0.01 (0.01)	-0.07 (0.01)	0.01 (0.01)	0.01 (0.01)
regular - small class (grade 1)	0.02 (0.02)	0.03 (0.03)		0.02 (0.02)
black - white	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)	-0.04 (0.03)
upper - lower SES	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)
girls - boys	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
regular - small class (kinder.)		0.05 (0.04)	0.00 (0.02)	
Random:				
Between schools:				
$\sigma^2_{(4)0}$	0.11 (0.02)	0.11 (0.02)	0.11 (0.02)	0.14 (0.03)
$\sigma_{(4)01}$	-0.02 (0.01)	-0.02 (0.01)	-0.02 (0.01)	-0.02 (0.01)
$\sigma^2_{(4)1}$	0.02 (0.004)	0.02 (0.004)	0.02 (0.004)	0.02 (0.004)
$\sigma^2_{(4)02}$				-0.03 (0.02)
$\sigma_{(4)12}$				0.00 (0.01)
$\sigma^2_{(4)2}$				0.04 (0.02)
Between classes:				
$\sigma^2_{(3)0}$	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)	0.06 90.01)
$\sigma_{(3)01}$	-0.03 (0.004)	-0.03 (0.004)	-0.03 (0.005)	-0.03 (0.004)
$\sigma^2_{(3)1}$	0.02 (0.003)	0.02 (0.003)	0.02 (0.003)	0.02 (0.003)
Between children:				
$\sigma^2_{(2)0}$	0.22 (0.01)	0.22 (0.01)	0.22 (0.01)	0.22 (0.01)
$\sigma_{(2)01}$	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)
$\sigma^2_{(2)1}$	0.01 (0.004)	0.01 (0.004)	0.01 (0.004)	0.01 (0.004)
Between years:				
$\sigma^2_{(1)0}$	0.19 (0.01)	0.19 (0.01)	0.19 90.01)	0.19 90.01)
$\sigma_{(1)01}$	0.01 (0.004)	0.01 (0.004)	0.01 (0.004)	0.01 (0.004)
-2*(log likelihood)	19210.5	19209.3	19216.3	19200.3

Number of students in analysis = 4182; number of measurements over all occasions = 11433; number of students with all four measurements = 2920.

Table 6b for Reading shows a between-school variation in the grade 1 class size effect, as in Table 5. It also shows a between-student variation in the trend over time, with students changing at different rates. The other results are similar to those for Mathematics except that neither the higher SES nor the black students show a gain after grade 1.

Further elaborations of these analyses are possible, including interactions among the variables and the introduction of teacher level variables, but we do not pursue these analyses any further

in this paper. Largely this is because the quality of the data, with the absence of any prior achievement measure and with differential dropout, does not seem to merit much further detailed exploration. Also, the main thrust of this paper is methodological and we have illustrated the major issues in our analyses of Tables 1 - 6.

13. In conclusion

Our examination of the methodology of class size studies can be summarised in two general conclusions. The first is that too little attention has been paid to the requirements for valid causal conclusions. These requirements include the need carefully to specify the reference population of interest, the need for good initial achievement data on students and the usefulness of measuring the *processes* occurring within classrooms including the expectations of teachers.

Secondly, it has often been assumed that randomised controlled trials are the only means of reaching causal type conclusions: our own analysis suggests that RCTs suffer from both practical and theoretical drawbacks which have received too little attention. Our reanalysis of the STAR study, the largest and most comprehensive RCT, has shown that it has limitations inherent in its initial design as well as its execution. Our analysis has attempted to compensate for the deficiencies associated with differential dropout and the lack of baseline measures. In particular, the use of multilevel modelling has shown that inferences about causality should not be made without studying the variation between schools. While we failed to detect any significant between-school variation in class size differences for Mathematics there were differences for Reading. This suggests that while our reanalysis provides corroboration for the conclusion of a modest but real class size effect, it may not be consistent across all schools nor for all measurements. More research is needed to further understand this issue.

Perhaps one of the most powerful arguments in favour of RCTs occurs when we wish to study new situations which do not occur naturally or not in sufficient numbers. This would be the case where we wished to study the effects of very small classes within a system where these did not exist, or were provided only for special groups of students such as those with learning difficulties. It is a common design for the evaluation of new educational or social initiatives and it is one of the standard situations for the application of RCTs in medicine, especially in the evaluation of novel drugs or treatments. On the other hand, it is difficult for RCT designs to simulate the reality of social systems, for example informative clustering of students, and this may severely limit the possibilities of generalising from the results of RCTs to the real world.

Observational studies of class size have also suffered from poor designs and inadequate analysis, but with careful attention to the requirements as we have outlined them, it should be possible for such studies to provide useful insights into the effects of class size and in particular to study the factors associated with differential effects across schools.

If we are to judge by the number of class size studies being carried out and the amount of political interest, this is an issue which will persist in importance. The limitations of existing work which we have pointed out have also encouraged us to see whether it is possible to improve considerably upon existing designs, and a new study has been started with this aim (Blatchford et al., 1996). This is an observational study with baseline measurements at entry to school and measures of class composition and change over a two year period. It is also collecting relevant teacher and school information and will utilise efficient multilevel modelling techniques for analysis. It will investigate the stability of class size effects across institutions and by type of student, especially in terms of initial baseline status. Its results, which will be reported elsewhere, should help further to enhance our understanding of the methodological issues.

Finally, we need to point out that our discussion has focused on establishing the minimum conditions which allow us to draw causal inferences from class size studies. We have said something about exploring the detailed *means* by which any change in class size actually produces changes in cognitive or affective attributes. There is, of course, no reason why a statistical modelling approach cannot be extended to studying such processes, although this would typically involve the collection of large amounts of detailed process data. To be effective, however, such research would benefit by being supplemented by detailed qualitative and case study research which can attempt to generate the specific theories for further evaluation and testing.

Acknowledgements

We are most grateful to the following who commented on drafts of this paper: Roel Bosker, John Gray, Michael Healy, Ann Oakley, Ian Plewis, Abby Riddell, Pam Sammons and John Smythe.

Abstract

The paper reviews research into class size effects from a methodological viewpoint, especially concentrating on the various strengths and weaknesses of randomised controlled trials and observational studies. It sets out the criteria for valid inferences from such studies and illustrates these using a reanalysis of the large data set from the Tennessee STAR study.

References

Akerhielm, K. (1995). Does class size matter? *Economics of education review* **14**: 229-41.

Bennett, N. (1996) Class size in primary schools in primary schools: perceptions of headteachers, chairs of governors, teachers and parents, *British Educational Research Journal*, 22,1,33-55

Blatchford, P and Mortimore, P. (1994). The issue of class size for young children in schools: what can we learn from research. *Oxford review of education* **20**: 411-28.

Blatchford, P., Burke, J., Farquhar, C., Plewis, I. and Tizard, B. (1987) 'A systematic observation study of children's behaviour at Infant school', *Research Papers in Education*, 2,1,47-62.

Blatchford, P., Mortimore, P. and Goldstein, H. (1996). *Class size and pupil's progress: a research proposal.* London, Institute of Education.

Cooper, H.M. (1989) Does reducing student-to-teacher ratios affect achievement? *Educational Psychologist*, 24,1, pp. 79-98

Creemers, B. (1994) *The Effective Classroom.* London: Cassell

Dunkin, M.J. and Biddle, B.J. (1974) *The Study of Teaching* New York: Holt, Reinhart & Winston

Glass, G. V. and Smith, M. L. (1979). Meta analysis of research on class size and achievement. *Educational evaluation and policy analysis.* **1**: 2-16.

Glass, G. V., Cahen, L. S., Smith, M. L. and Filby, N. N. (1982). *School Class Size*. Beverly Hills, Sage.

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

Gray, J. and Wilcox, B. (1995). *Good School, Bad School*. Buckingham, Open University Press.

Hedges, L. V. and Olkin, I. O. (1985). *Statistical methods for meta analysis*. Orlando, Florida, Academic Press.

Hill, P. W. and Goldstein, H. (1997). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural Statistics, (to appear).*

Jamison, D. T. (1987). *Reduced class size and other alternatives for improving schools: an economists view.* Washington, DC, World Bank.

Johnston, J. (1972). *Econometric methods*. New York, McGraw Hill.

Mitchell, D. E., Beach, S. A. and Badaruk, G. (1991). *Modelling the relationship between achievement and class size: a re analysis of the Tennessee project STAR data.* Riverside, California, California Educational research co-operative.

NAHT (1996). *Class size research and the quality of education*. Sussex, National Association of Head Teachers.

Nye, B. A., Achilles, C. A., Zaharias, J. B., Fulton, B. D., et al. (1993). Tennessee's bold experiment: using research to inform policy and practice. *Tennessee Education* **23**: 10-17.

Preece, P. F. W. (1987). Class size and learning: a theoretical model. *Journal of Educational research* **80**: 377-79.

Robinson, W. S. (1951). Ecological correlations and the behaviour of individuals. *American Sociological Review* **15**: 351-7.

Shapson, S. M., Wright, E. N., Eason, G. and Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal* **17**: 144-52.

Slavin, R. (1990). Class size and student achievement: is smaller better? *Contemporary Education* **62**: 6-12.

Wilmms, J.D. (1992) *Monitoring School Performance*: *A Guide for Educators.* London: Falmer.

Word, E. R., Johnston, J., Bain, H. P., Fulton, B. D., et al. (1990). *The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report 1985-90.* Nashville, Tennessee State University.