Class Size and Educational Achievement: a review of methodology with particular reference to study design[1]

HARVEY GOLDSTEIN & PETER BLATCHFORD, *Institute of Education, University of London*

ABSTRACT *The article 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 discusses population definitions, causation and generally sets out the criteria for valid inferences from such studies. For illustration it presents some new findings from a reanalysis of the large data set from the Tennessee STAR (Student Teacher Achievement Ratio) study.*

Introduction

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 article 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 a review of existing findings see Blatchford & Mortimore (1994) and Slavin (1990).

In the course of this article we will explore the methodology which has been used to date. We will look at both observational [2] studies and randomised controlled trials (RCTs) and will endeavour to establish criteria for judging the usefulness of different study designs.

In particular we shall report part of a reanalysis of the Student Teacher Achievement Ratio (STAR) project data; a large ($11 million) RCT in Tennessee from the 1980s. Since studies of class size take place within existing educational systems which are organised into complex hierarchical structures, with students being grouped within class-

Received: 8 October 1997; resubmitted: 17 December 1997; accepted: 18 December 1997.

0141-1926/98/030255-14 ©1998 British Educational Research Association

rooms and the latter grouped within schools, it is appropriate to use multilevel statistical models in the analysis of such data, but we will not go into details about these here.

Underlying the discussion is the assumption that the point of doing class size research is to make statements about causation. By causation we mean the inference that, from an observed 'effect' of class size on achievement 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; and 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. In an observational study it is necessary to make such an adjustment 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 and develop 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 RCTs to studies of class size. We then look at issues of causality and factors which may explain the effect of class size upon attainment. In the light of this review we describe some reanalysis of the STAR data for reading and mathematics achievement. Finally, we shall draw some conclusions for future research.

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 multigrade 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.

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 later. 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: 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. 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 an 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 cooperation 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 *et al.*, 1993) schools were required to agree to participate in the study for 4 years and had to supply any extra accommodation necessary. 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.

Randomised Controlled Trials

Randomisation of subjects to different 'treatments' or experimental situations typically 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 inferences are correct, on average, to all populations of interest; for example, with particular proportions of classes with high percentages of low achievers.

To see this, consider the extreme case where the only difference between small and large classes is where the percentage of low achievers is more than, say, 50%. Suppose the average proportion of low achievers is 10%. If random allocation has taken place without stratifying for the proportion of low achievers, then for typical classes of size 15–25 the probability of a class having at least 50% low achievers is extremely small, less than 1 in a thousand; and, importantly, being less the larger the class. Thus, even in reasonably sized studies it would be unusual to find classes with high proportions of low achievers. Even where studies did have such classes there would tend to be more of them where the class size was larger. In the first case we would be unaware of any effect, even though in the real population classes with high proportions of low achievers did exist. In the second case we would obtain an estimate of the average difference, due to class size, only for a population with the same distribution of this compositional variable as found in the study. To avoid this difficulty, we would need to adjust for the compositional factor. This means that we are required to explore statistical models which adjust for relevant factors in order to arrive at valid causal inferences and these explorations are, formally, the same as those used in observational studies. We see here, therefore, an instance of where a key rationale for randomisation, namely the equalisation (on average) of initial characteristics within the 'treatments' being studied, undermines the possibility of valid inferences.

Those designs where randomisation is within schools face particular problems. 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 an 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, and 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 difficulty 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. 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 *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. It is also possible that in an RCT some teachers of large classes may work harder than their colleagues in order to compensate for a prior expectation that large classes are less effective. If this occurs it may induce between-school variability in any class size effects.

A design such as STAR, where each school has one or more small classes and one or more 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 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 assign units randomly, 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 are actively involved in making autonomous decisions.

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 (adjusted) 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 to 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 are 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.

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 USA, and doubts exist about the reliability of some of the studies (see Blatchford & 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 & 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 appear to have little effect. In order to examine these possibilities more closely, detailed information on classroom processes would be needed.

Although we shall not review the research on mediating factors (for reviews see Cooper, 1989; Blatchford & Mortimore, 1994; National Association of Head Teachers [NAHT], 1996), some relevant methodological issues can be identified. 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. Pupils' 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 & 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 & 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.

Non-cognitive Responses

The discussion so far 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 (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.

Cost–Benefit Analysis

The principal focus of this article 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.

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 article we shall present a reanalysis of some of the data from that study in order to illustrate some of the methodological points we have been making. These results are extracted from a much more extensive reanalysis reported in Goldstein & Blatchford (1997). Although there have been other critical commentaries on the STAR study (for example, Mitchell *et al.*, 1991; Prais, 1996), these have not undertaken reanalyses of the original data.

For present purposes we will look only at mathematics and reading achievement throughout the 4 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.

The full reanalysis looked at a small number of key explanatory variables. It explored the data through a series of models of increasing complexity in order to illustrate ways in which the use of multilevel modelling techniques can uncover relationships and test causal hypotheses (Goldstein, 1995). Here, we use only a summary of the results from the kindergarten and grade 1 stages.

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 *et al.*, 1990) notes that this was to 'achieve sexual and racial balance and to separate incompatible children'. Table I 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. Subsequently the regular and regular with aide class types are amalgamated since they exhibit few differences. There is an overall difference in favour of the small classes, whether classified by kindergarten membership (0.16 units for mathematics and 0.17 for reading) or grade 1 membership (0.29 for mathematics and 0.26 for reading). For mathematics those who were in small kindergarten classes had a lower kindergarten score if they moved to a regular class in grade 1 (by 0.24 units), with a smaller decrease for those in regular kindergarten classes who moved to small grade 1 classes (of 0.08 units). Similarly, 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 (0.39 and 0.07 respectively). Generally, the results are similar for mathematics and reading. There were 24% who were lost to the study after kindergarten, and these had a markedly lower score than those who remained in the study. 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. Note also that a higher proportion of those in regular kindergarten classes were lost to the study than those in small classes and it is not clear why this occurred. Such a differential loss may explain some of the subsequent findings about the relative lack

TABLE I. Mean mathematics and reading score at the end of kindergarten for kindergarten by grade 1 class type. Numbers of children in brackets. Scores are standardised to have zero mean and standard deviation 1

Kindergarten	Grade 1			
	Small	Regular	Missing	Total
Mathematics				
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)

of further differences between class sizes following the grade 1 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. 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.

The conclusion from Table I about class size is that there is a difference in favour of the small classes of about 0.15 standardised units (standard deviations) and that this is the same at the end of kindergarten and the end of grade 1. The remaining analysis looks at the mathematics and reading attainment at the end of grade 1.

Achievement at the End of Grade 1

Table II shows some selected results illustrating different interpretations from different statistical analyses. These are derived from a more extensive, multilevel analysis where significance tests have also been carried out (further details are in Goldstein & Blatchford, 1997).

For mathematics the unadjusted class size differences for black and white children are larger than those at the end of kindergarten, with that for black children being somewhat greater than that for white children. When end-of-kindergarten attainment is allowed for there are still differences indicating a further effect of class size in grade 1, especially for black children. For reading we have a similar picture, but now after adjustment there is no additional effect for white children in grade 1.

For reading, the negligible effect for white children, with a still substantial effect for black children, may have important policy implications. This effect emerges only after kindergarten since there is no apparent 'interaction' between class size and race when the end-of-kindergarten score is chosen as the outcome (Goldstein & Blatchford, 1997). If this result is accepted as 'causal' then it suggests a policy of allocating black children, and perhaps disadvantaged children more generally, to smaller classes following the first kindergarten year.

A particularly interesting result is the between-school standard deviation of 0.25 for reading. This is of the same order of magnitude as the overall, adjusted, class size difference and implies that the 'effect' is negligible in some schools (or even reversed) and very large in others. This variability may be related to the problems of a zero-blind RCT, especially those associated with teacher expectations, which we discussed earlier.

TABLE II. Selected (standardised) effects for mathematics and reading at end of grade 1; with and without adjusting for end-of-kindergarten score

	No adjustment	Adjusted
Mathematics		
Small–regular: Black	0.40	0.35
White	0.29	0.18
Reading		
Small–regular: Black	0.32	0.21
White	0.13	0.04
Between-school standard deviation of class size difference (based upon a multilevel model)		0.25

If the level of awareness of the experiment and the responses of the teachers differed among schools, we might well expect to see a variation in the class size difference.

Beyond grade 1 the problem of differential drop-out persists. Thus, of those with data at grade 1, 21% have no data on mathematics or reading at grade 2, and of those with data at grade 2, 14% have none at grade 3. Those who dropped out after grade 1 have particularly low scores (0.32 units below average for mathematics and 0.25 units for reading). Further analyses for grades 1–3 have been carried out by Goldstein & Blatchford (1997) and show some further, but rather small and variable changes after grade 1 in relation to class size.

In Conclusion

This examination of the methodology of class size studies can be summarised in two general conclusions. The first is that attention has to be paid to the requirements for valid causal conclusions. These requirements include the need to specify carefully 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 RCTS are the only means of reaching causal type conclusions: the present article suggests that RCTs suffer from both practical and theoretical drawbacks which have received too little attention. 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.

In substantive terms the use of multilevel modelling has indicated some variation in the class size effect among schools for reading. The analysis has also suggested that after the first kindergarten year the 'effect' of class size on progress may be more important for black children than for white, and this could have far-reaching policy implications.

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 2-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 this discussion has focused on establishing the minimum conditions which allow us to draw causal inferences from class size studies. Less has been said 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.

Correspondence: Harvey Goldstein and Peter Blatchford, Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK; e-mail: h.goldstein@ioe.ac.uk; p.blatchford@ioe.ac.uk.

Acknowledgement

We are most grateful to the following who commented on drafts of this paper: Chuck Achilles, Roel Bosker, Jeremy Finn, John Gray, Michael Healy, Barbara Nye, Ann Oakley, Ian Plewis, Abby Riddell, Pam Sammons and John Smythe. We are most grateful to the Research and Policy Centre at Tennessee State University for its support and for access to the STAR data. This paper was originally prepared for UNESCO and we are most grateful to that organisation for its encouragement and support.

NOTES

[1] This is based upon a longer paper by Goldstein & Blatchford (1997).

[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.

REFERENCES

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

Blatchford, P. & Mortimore, P. (1994) The issue of class size for young children in schools: what can we learn from research. *Oxford Review of Education*, 20, pp. 411–428.

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

Blatchford, P., Mortimore, P. & Goldstein, H. (1996) *Class Size and Pupils' Progress: a research proposal* (London, Institute of Education).

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

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

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

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

Goldstein, H. & Blatchford, P. (1997) Class size and educational achievement: a methodological review, paper prepared for UNESCO (London, Institute of Education). (This document may be downloaded from http://www.ioe.ac.uk/hgoldstn/csize-download.html)

JAMISON, D. T. (1987) *Reduced Class Size and other Alternatives for Improving Schools: an economist's view* (Washington, DC, World Bank).

MITCHELL, D. E., BEACH, S. A. & BADARUK, G. (1991) *Modelling the Relationship between Achievement and Class Size: a reanalysis of the Tennessee project STAR data* (Riverside, CA, California Educational Research Co-operative).

NATIONAL ASSOCIATION OF HEAD TEACHERS (1996) *Class Size Research and the Quality of Education* (Haywards Heath, 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, pp. 10–17.

PRAIS, S. J. (1996) Class size and learning: the Tennessee experiment—what follows? *Oxford Review of Education*, 22, pp. 399–414.

SHAPSON, S. M., WRIGHT, E. N., EASON, G. & FITZGERALD, J. (1980) An experimental study of the effects of class size, *American Educational Research Journal*, 17, pp. 144–152.

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

WILLMS, J. D. (1992) *Monitoring School Performance: a guide for educators* (London, Falmer Press).

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, TN, Tennessee State University).