Class size and educational achievement: new evidence?

(*Revision - July 11, 2000)*

Current educational policy in the UK and elsewhere has emphasised initiatives to reduce class sizes. To a large extent these policies derive from recent research and research reviews that have demonstrated moderate learning gains from class size reductions. Any new research, therefore, that appears to contradict this deserves careful attention and evaluation. One such recent study is that of Hoxby (2000) who presents an analysis of data on class size which concludes that 'reductions in class size have no effect on student achievement'. Such a categorical statement clearly has important implications, if true, and apart from anything else appears to contradict most existing research findings. In this review I evaluate the research that underpins these claims and similar ones by Rivkin, Hanushek and Kain. (2000), and conclude that various methodological weaknesses render such conclusions unsound.

1. The existing evidence:

Class size is one of the most studied issues in educational research Glass (1979) and later Slavin (1990) carried out extensive reviews of the literature, and there have been several subsequent commentaries too (See e.g. Blatchford et al., 1998 and Goldstein and Blatchford 1998). These commentators are in agreement that most reported studies are flawed, either because they failed to make adjustments for key factors, such as prior achievement in purely cross sectional observational studies or because of sampling inadequacies. In his meta analysis Slavin (1990) selects only 9 studies as meeting strict criteria for safe inferences and Yang et al., (2000) in their meta analysis come to a similar number and also demonstrate that there is considerable agreement in the magnitude of class size effects, namely 0.2 - 0.3 of a standard deviation improvement in test score associated with a reduction of 10 pupils, in the very early years of schooling; this effect size being comparable for both randomised controlled trials and observational studies.

The recent study most often cited in the literature is the Tennessee STAR study (Word et al., 1990) whose results are in line with those above. This was a randomised controlled trial (RCT) of some 6,000 pupils and some 70 schools which also found that black children appeared to benefit more than white children. While the methodology of this study has been questioned (Goldstein and Blatchford, 1998) together with the other studies alluded to above, it does support the case for a moderate class size effect in the early years.

Hoxby's paper mentions the STAR study, but emphasises its deficiencies as a prelude to her own study which claims to find little effect. She completely fails to acknowledge any of the other literature reviewing class size, including that of other North American academics such as Glass and Slavin,. This failure to cite the relevant literature does weaken any case she makes for statements about class size effects.

2. Hoxby's data and statistical models

The data come from a long time series of enrolment into a cohort starting in kindergarten. She makes the assumption that the nature of the population supplying the schools undergoes only gradual change in size and composition that can be smoothly modelled over time. Because of random fluctuations in births, it is argued, class sizes will change between cohorts, and occasionally because of school board policies on maximum class sizes will change markedly if a maximum threshold is exceeded, resulting in two much smaller classes. While

Hoxby does discuss complications due to voluntary migration affecting population composition (in terms of social status), there is no evidence provided that such factors have not influenced the 'steady state' assumption of the population here. Likewise there are no data available to study the possible effect of differential teacher allocation strategies when class sizes vary from cohort to cohort. The only reliable procedure that would adjust for such unobservable factors is to collect data on pupil achievement at the start of kindergarten (or other grade) and use this an adjustment variable in a subsequent model.

Such a procedure, however, is only possible if the analysis is carried out at the individual student level, taking into account the higher levels of school and district in a full multilevel formulation. In Hoxby's analysis, however, the unit of analysis is the school (for each cohort) - not even the class within school, so this possibility is unavailable. Furthermore, cohort analysis of this kind is unable to take account of within-cohort mobility, so that while test data are nominally available for the 'same' cohort this will typically contain different pupils over time. We are given no information on the extent of such mobility or the causes of it.

In fact, Hoxby's analysis is in real danger of committing the 'ecological fallacy' - namely assuming that relationships that hold at a higher level of aggregation, that is the school, also apply to individual children.

Hoxby claims to find little association between class size and achievement at the level of the school, and even leaving aside the problems discussed above this says little about the possible relationship at the pupil level when individual pupil characteristics such as prior achievement, mobility and social background are allowed for. In essence, Hoxby's analysis is indirect, relying upon some untestable assumptions. It also suffers from the fact that her cohorts are successive cohorts within each school. The fact of class size varying between these is the raison d'etre for the analysis, yet the very existence of such variation may trigger within school mechanisms that compensate for the perceived 'disadvantage' of a large class. This possibility is not raised in the paper.

In short, the analysis that is presented is flawed and the conclusions derived from it need to be treated with extreme caution.

Like Hoxby, Rivkin et al. use aggregate data for each school in his large Texas sample. Like Hoxby they also use gain scores - in two analyses. The first analysis attempts to find a lower bound for the influence of teachers. Unfortunately, the statistical model is not only at a high level of aggregation, it also ignores any dependence of the gain, or difference between successive gains, made by a student, on the prior achievement of that student. Thus, the claim that this model does indeed allow for individual student effects is untrue and the existing literature demonstrates is quite clear that standardised gain scores are related to the prior achievement component of such a score (see e.g. Yang et al., 2000). The failure to include this means that this component will form part of the estimate of the between-teacher variation. Thus, their claim that this variation is *large* needs to be read with caution. While there may indeed be large amounts of teacher variation, it is not demonstrated by this analysis. Likewise, in the second analysis concerned with class size effects the model again ignores the prior achievement of the students and is carried out at the school level.

Neither Rivkin et al. nor Hoxby seem to be aware of the large literature on multilevel modelling of student teacher and school effects which uses sophisticated models based upon individual student data and includes random coefficient models as well as basic variance component ones (see e.g. Yang et al, 2000). In the case of the Rivkin et al. paper, the failure to use individual student data, which does in fact seem to be available is a pity, since this would provide a much more useful analysis.

My conclusion, therefore, is that both these reports, while claiming to provide important new evidence about class size and teacher effectiveness, fall short of acceptable standards for valid inference and contribute little useful evidence to the debate.

It is also worth remarking, in the context of current interest in evidence based policy making, that on this particular issue, as I have attempted to demonstrate, a comprehensive review of the area necessarily requires a detailed evaluation of the statistical techniques employed. More generally, this has important implications for how systematic research reviews are implemented. At the very least there needs to be a procedure for understanding when such detailed technical expertise is required involving individuals who themselves possess such expertise.

References

Blatchford, P., Goldstein, H. and Mortimore, P. (1998). Research on class size effects: a critique of methods and a way forward. *International Journal of Educational Research* **29**: 691-710.

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.

Goldstein, H. and Blatchford, P. (1998). Class size and educational achievement: a review of methodology with particular reference to study design. *British Educational Research Journal* **24**: 255-268.

Hoxby, C. M. (1998). *The effects of class size and composition on student achievement: new evidence from natural variation*, National Bureau of Economic Research.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2000). *Teachers, schools and academic achievement*, National Bureau of Economic Research.

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

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.

Yang, M., Goldstein, H., Omar, R., Turner, R., and Thompson, S. G. (2000). Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, Series C,* **49**: 1-14.

Harvey Goldstein

July 2000