Oxford Review of Education, Vol. 10, No. 1, 1984

The Methodology of School Comparisons

HARVEY GOLDSTEIN

This article explores the assumptions which underpin research studies of the effects of pupil, school and Local Education Authority (LEA) characteristics on pupil achievement. While not intended as a critique of particular studies, it will have something to say about a number. Its aim is to clarify the necessary limitations on knowledge imposed by various research methods and to suggest how these might be improved. The first part considers individual or pupil-based studies and the second part tackles the problem of how to analyse and interpret data at higher levels of aggregation, such as the school or the LEA.

SCHOOL DIFFERENCES

In the study of the 'outcomes' of schooling, we are forced to make comparisons between schools based on schools as they exist rather than as they might be in, say, an experimental programme. Thus, for example, if we wished to determine whether smaller schools resulted in higher mathematics attainment, an experimental study would assign children at random to different size schools. These schools would differ only in their size so that any subsequent differences in terms of maths attainment could be attributed to that factor. In real life, typically, we cannot randomly assign children to schools, nor ensure that schools differ on only a single factor such as their size, and we have to search for alternative approaches.

First some obvious and substantial school differences do exist. Children in grammar schools, for example, on average achieve better exam results and score higher on achievement tests than children in secondary modern or comprehensive schools. The reasons for this are fairly clear; namely that grammar schools tend to select pupils who already have high achievements at the time of secondary school transfer. Thus, any outcome differences may simply be reflections of such intake differences and not attributable to the type of school *per se*.

Thus it would be more sensible to ask the following question. For a group of pupils of the same attainment immediately prior to entering secondary school, do those who go to grammar or secondary modern schools achieve higher or lower average exam results than those who go to comprehensive schools? We might also want to know whether such average differences were constant across different groups of children, for example, for each social class or for boys as well as girls. We would also want to know whether differences were the same for each pre-existing attainment or whether, for example, a grammar–comprehensive school difference decreased with increasing intake attainment, as the National Children's Bureau (NCB) National Child Development study (NCDS) found in some of its analyses.

A study which failed to compare school types on the basis of given intake attainment would face extremely difficult, if not insurmountable, problems in

attributing any causal interpretation to its findings. Nevertheless, even where such intake 'allowances' can be made adequately, there still remain considerable difficulties which the remainder of this article will consider.

SCHOOL DEFINITIONS

An immediate issue confronting research into differences between selective and non-selective schools is to be able to say clearly what is a 'true' grammar, secondary modern or comprehensive school. If we are interested in inferring the consequences of moving, say, from a fully selective to a fully comprehensive system, then there need to be some fully comprehensive or some fully selective systems or LEAs available for comparison. If we wish to infer the consequences of moving from one type of mixed comprehensive/selective system to another then likewise examples of these must be available.

Interestingly, in discussions of this issue, most research has been concerned essentially with inferences about differences between 'pure' systems, although there are many useful inferences which can be based on comparisons of 'mixed' systems. In England and Wales there are some fully comprehensive LEAs, some fully selective and some mixed. It is perfectly reasonable to make comparisons between different school types for each of the different systems. Thus, a comparison between grammar schools and also between secondary modern schools in fully selective systems with those in partly selective or mixed systems (some selective and some comprehensive schools) is both feasible and likely to be informative. Likewise, there is merit in a comparison of comprehensives in mixed and pure systems. Yet such comparisons seem not to have been attempted.

The NCB study did seek at one stage (Steedman, 1980) to introduce a 'degree of LEA selectivity' variable but did not attempt to use it as suggested above and abandoned it in its main comparisons. The importance of measuring 'selectivity' adequately when carrying out comparative analyses is due to what is often referred to as 'creaming', whereby the existence of selective schools within an LEA effectively prevents comprehensives capturing the full achievement range. There are different ways of measuring and taking account of this factor, but it is sufficient to note that the term 'comprehensive' needs to be qualified by the degree of comprehensiveness which is present. Moreover, there will be variations between comprehensives in this respect even within fully comprehensive LEAs, where some schools may look more like selective schools than others. Thus, in order to make useful inferences, careful measurement of school characteristics is essential.

MEASUREMENTS

There are many difficulties and there is perhaps not much agreement on how to measure academic outcomes of secondary schooling. While they are important, I do not wish to discuss at length the issues here but rather to assume that some decision will have been made. The NCB studies, for example, adopted an eclectic approach, using test scores, behaviour measurements and exam results. The latter currently seems to be the main focus of interest, partly no doubt because of the obligation of secondary schools under the 1980 Education Act to make public their examination results. In view of this it is worth remembering the considerable difficulties posed by such a measure. There are problems of scaling grades and assuring some comparability of grading schemes between exam boards or over time. These issues are still

largely unresolved (see, e.g. Goldstein, 1982). There are also considerable problems in deciding upon suitable measures of intake achievement. For example, if we were to take as a measure of outcome a reading comprehension test score, it would seem reasonable to use a reading comprehension test to allow or adjust for intake achievement. If, however, we are interested in exam results in 'O' level history, then it is by no means clear what kind of intake measure should be used. This problem is even more serious when an average over several examination subjects is used. In practice, for example in the Inner London Education Authority (ILEA, 1980), verbal reasoning tests typically have been used. A main justification for this would be if the verbal reasoning tests were used to allocate children to the ILEA schools, because one would then be adjusting for the factor on which intake differences depended. In general, however, and presumably for the ILEA too, there is no single measure which determines school allocation—rather a mixture of academic, social, geographical and other factors.

If we cannot measure such factors adequately, then causal inferences are jeopardised. This will also be true, for example, in the reading test case, since there may well be factors other than performance on an 'intake' reading test which influence later reading performance. Attempts to allow for such effects, while they need to be made, may not eventually succeed, and this is an example of the general difficulty of making causal inferences from purely observational studies.

TYPES OF PUPILS

Where individual pupils are the units of the analysis and intake measures are available, then an appropriate analysis is one where school types are compared on outcome measures, after making allowance for intake differences. In the simple case of an intake test score and a similar outcome test score, an 'analysis of covariance' or 'pre-post test' analysis is commonly done. Such an analysis gives estimates of the average school outcome score for each type of school for each intake score. (Where the intake measurement has a low reliability, special care needs to be taken with the analysis—Goldstein, 1979.) Where the 'adjusted' mean scores for school types differ, then we may infer that the average progress in schools of different types also differs.

Such analyses were carried out by the NCB. In these, the grammar, secondary modern and comprehensive school differences changed with intake attainment score. Thus, in the middle range of attainment in reading and mathematics, the mean outcome score for the grammar schools was higher than that for the comprehensives, but this shrank to a negligible difference in the top 20% of the attainment range. Such 'interactions' between attainment and school type provide the kind of detailed information which is far more valuable than statements only about overall school differences, because they begin to indicate how such differences may have arisen. Moreover, such analyses clearly can be done only if measurements are carried out and analysed for individual pupils.

UNITS OF ANALYSIS

A key to understanding what can be learnt from studies of educational processes lies in a clear conceptualisation of the different levels of organisation of the educational system.

In theory as well as in practice, most educational systems are hierarchical. Pupils are grouped in classes, which are grouped in schools which are grouped within

education authorities. At each of these levels there are factors operating which can affect the outcomes of schooling. Thus, pupils bring their own individual cultural attitudes, knowledge, etc. Classroom teachers bring their own teaching styles and experiences. Schools contribute forms of organisation and education authorities contribute resources. At any one level, all the units at that level share the same set of characteristics of the unit at the next higher level within which they are grouped. Thus, all the schools within an authority with a high proportion of middle-class households will be classified as belonging to a high middle-class area. Likewise, at all levels except the pupil level, units can be classified by the characteristics of lower level units. A class can be classified by the average test scores of the pupils in it, or, for example, by the standard deviation of those test scores.

PUPILS AS UNITS

I have argued that analyses at the pupil level are useful because they make it possible to study the effects of children's differing characteristics. Important as this is, however, it presents an incomplete picture unless characteristics of higher level units are incorporated into the analysis. Thus, the social background of the school, its attainment range, the characteristics of its teachers and the policies of the LEA are potentially important influences on pupil achievements. If such information is to be obtained, then a study needs to be designed hierarchically, perhaps using a multi-stage random sample which first selects a sample of LEAs, then a sample of schools within these LEAs, classes within schools and finally pupils within classes.

With such a design, the characteristics of units at each level can be measured readily. On the other hand, if a random sample of children is selected from the total population of children, which effectively is the case with the NCDS, then it becomes more difficult to measure the characteristics of higher level units. Thus the NCDS never has more than a handful of children belonging to any one class, so that using these data alone, it was not possible to obtain good data about the variability of any school's attainment, or at LEA level details of how selection policies operate. Such data are better and more easily obtained using a multistage procedure.

Thus, despite providing currently the best available data for comparing school types, because it is individually based, and is longitudinal, the NCDS is deficient in these other respects. Of particular importance are measures of LEA selection procedures and how their operation affects individual schools. Thus, as already pointed out, comprehensive schools vary widely in terms of their attainment distributions and in order to make full allowance for these, we would need to collect full attainment data from schools. Likewise, the LEA selection and school allocation procedures will help to determine the nature of different schools.

AGGREGATE LEVEL ANALYSES

A common substitute for individual level analyses is to analyse at the level of schools or LEAs. Thus the Inner London Education Authority (ILEA, 1980) carried out a longitudinal study of average school examination results using average school intake test scores. The National Council for Educational Standards (NCES) (Marks, Cox & Pomian-Srzednicki, 1983) used average school examination results in a study which was not longitudinal and used an average LEA social class measure as an 'adjustment' variable.

It is well known and anyway fairly obvious, that relationships at, say, the

individual level may be quite different from the relationships which exist between the same variables measured at the school level. Thus there may be differences in rates of progress between different types of school using pupils but not using schools as units. In the absence of suitable data at all levels, therefore, we are not able to make, in general, useful inferences about lower level relationships from only higher lever data. The question, therefore, is how much use such higher level data on its own might have.

Consider, by way of illustration, the case of school level data within an LEA. Suppose there are measures of average school attainment at intake and for outcome, but no pupil level data. Then we can compare, say, grammar and comprehensive schools in terms of their average attainment progress. We can study school factors associated with any differences, but we cannot know what happens to different kinds of pupils within the schools. Moreover, there is evidence from the ILEA analyses that at least in the case of school examination results, these are highly predictable from simple measures of average school intake attainment and social class—correlations of more than 0.90 readily being attainable. This suggests that there may be few other characteristics of schools which are important, and if so it may simply mean that in terms of average characteristics, schools have little effect, despite possibly large effects at the individual pupil level. Thus, if analyses are confined to the school level, we may rapidly reach the position when everything of any importance has been explained by a few simple factors and further analysis is relatively uninteresting. A similar argument will apply if other school level statistics are studied; for example the attainment variance or the proportions in specified attainment groupings.

The argument, with even greater force, applies to LEAs as units of analysis. Thus, analyses relating LEA expenditure to, say, examination results such as carried out by NCES and the Department of Education and Science (DES), are very likely to yield negligible relationships, not because pupil attainment is unaffected by expenditure, but because, having adjusted for social and other variables, little is left to explain at the aggregate level. Also, in this case, there is the difficult problem that there are relatively few LEAs and hence we are relatively unlikely to detect other than gross differences.

One further, and perhaps crucial, problem remains. In the ILEA analyses, it is apparent that there is a group of schools with high average intake scores, these being selective grammar-type schools which still existed in the mid-1970s. If we wished to use such data to compare types of school, there would in fact be too little overlap between these and comprehensive schools to make meaningful comparisons over a useful range of intake attainment scores. Indeed, it would be surprising if there was much overlap at all within an LEA since this would imply that there were some comprehensives with higher average intake test scores than some grammar schools and at the other end of the range with lower average intake scores than some secondary modern schools.

To some extent the same problem exists at the pupil level, but here the selection process is not completely reliable. For example, tests and teacher ratings are subject to error and geographical factors may contribute to less than a complete separation between school types in terms of intake attainment. Thus, typically there will be a useful range over which school types can be compared, although at the extremes of the ability range this may well not be possible.

The NCDS was able to make comparisons over a wide range of attainment because its sample was spread over all LEAs including fully comprehensive and fully

selective ones. In terms of school averages, however, one still might not expect very much overlap. If much overlap did exist for a study, say, of one fully comprehensive and one fully selective LEA, then for similar reasons to those given above, one might want to query how 'comprehensive' the former LEA really was.

When we come to consider analyses, for example, of LEA expenditure in relation to examination results with adjustment for intake attainment, the above arguments do not apply with the same force. If the LEA is the unit of analysis then expenditure presumably is not associated with intake attainment in the same way as is type of school in a school level analysis. Nevertheless, the other drawbacks of LEA level analysis which have been discussed are still present.

CONCLUSIONS

In conclusion, if the arguments of this paper are accepted, there are deficiencies in all existing comparative studies of school outcomes. The individually based NCDS, while providing the best available data for selective and non-selective school type comparisons, does lack important information on school and LEA effects and the school and LEA based analyses of the DES and NCES lack the all-important pupil level data.

In arguing for a full multistage or multilevel study design, I do not wish to claim that designs which analyse only at school or LEA level are of no use, merely that their uses are limited. Certainly such studies are important in administration and to monitor overall trends in the educational system. Nevertheless, despite their research drawbacks, there is scope for a careful exploration of the precise location of the limitations of such studies, especially in the area of school examination results, where there is now a danger that they will be used in ways which may not be appropriate. Finally, one may ask where this leaves us in terms of school type comparisons. Perhaps the most useful summing-up of the current position was given in the letter discussing the NCB report, written to the Education Secretary in 1982 by the advisory group appointed by the DES to oversee this study:

> the evidence rules out both the hopes of those who advocated comprehensives as the panacea for all educational ills and the fears of those who opposed them as disastrous 'social engineering'. The record of the comprehensives as they were in reality in 1974 is much the same as that of other state schools—neither a triumph nor a defeat.

REFERENCES

GOLDSTEIN, H. (1979) Some models for analysing longitudinal data on educational attainment, *Journal of the Royal Statistics Society A*, 142, pp. 407–442.

GOLDSTEIN, H. (1982) Models for equating test scores and for studying the comparability of public examinations, *Educational Analysis*, 4, pp. 107–118.

INNER LONDON EDUCATION AUTHORITY (1980) *School Examination Results in the ILEA 1978.*

MARKS, J., COX, C. & POMIAN-SRZEDNICKI, M. (1983) *Standards in English Schools* (London, National Council for Educational Standards).

STEEDMAN, J. (1980) *Progress in Secondary Schools,* (London, National Children's Bureau).

STEEDMAN, J. (1983) *Examination Results in Selective and Non-Selective Schools* (London, National Children's Bureau).

References

Some Models for Analysing Longitudinal Data on Educational Attainment

Harvey Goldstein

Journal of the Royal Statistical Society. Series A (General), Vol. 142, No. 4. (1979), pp. 407-442.

Stable URL:

http://links.jstor.org/sici?sici=0035-9238%281979%29142%3A4%3C407%3ASMFALD%3E2.0.CO%3B2-0