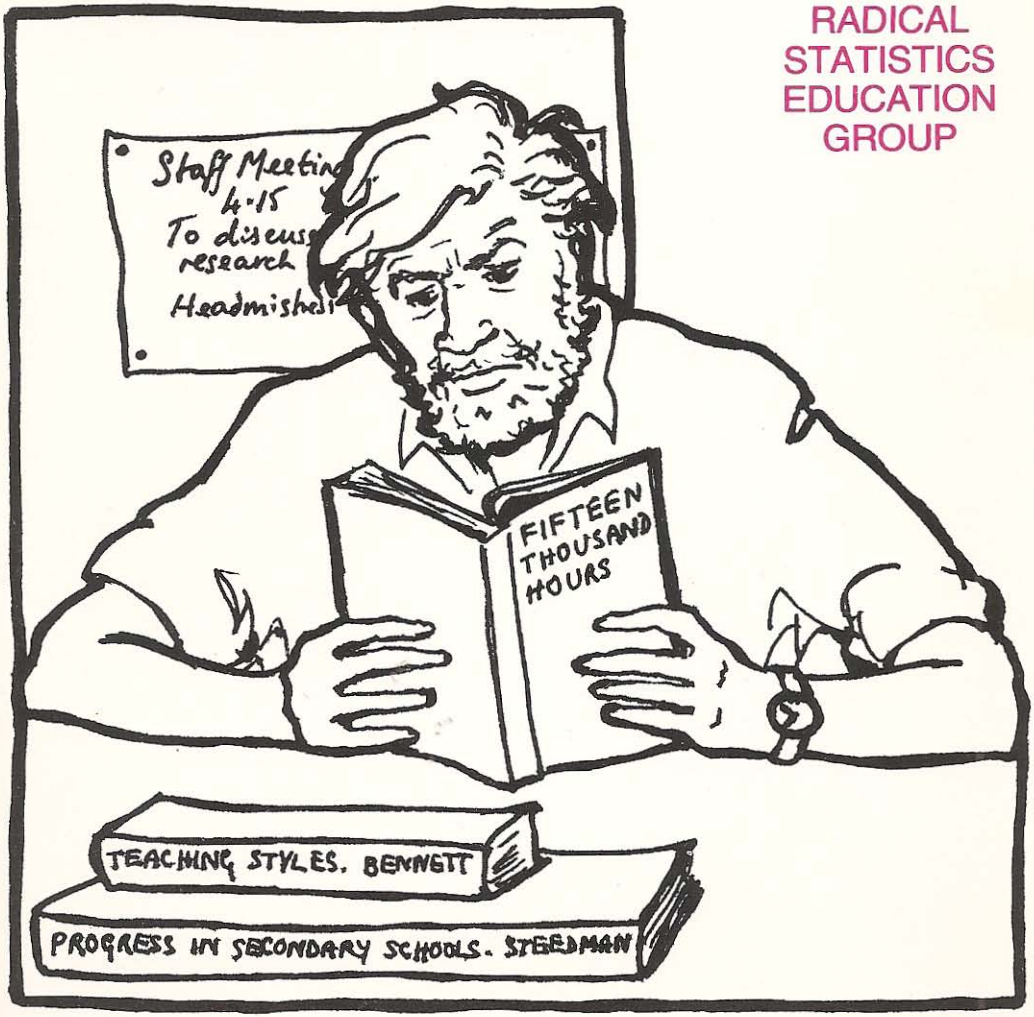
READING BETWEEN THE NUMBERS

A CRITICAL GUIDE TO EDUCATIONAL RESEARCH



RADICAL STATISTICS EDUCATION GROUP

RADICAL STATISTICS EDUCATION GROUP

READING BETWEEN THE NUMBERS

A CRITICAL GUIDE TO EDUCATIONAL RESEARCH

BSSRS Publications Ltd.

This pamphlet was written by the Radical Statistics Education Group, which includes statisticians and social scientists who are involved in teaching and research using statistics.

Contributors include: Russell Ecob, Jeff Evans, Dougal Hutchison, Ian Plewis.

For extensive written contributions, thanks are due to Tony Fielding, Harvey Goldstein, Helen Quigley and Ludi Simpson.

We should also like to thank the many people who attended and contributed to the early discussions of the group from which this pamphlet took shape, who provided critical comments on early drafts, or who helped with production, particularly typing: Brian Clover, Dick Cooper, Ken Fogelman, Pamela Glanville, Elizabeth Goodacre, John Gray, David Hamilton, Martyn Hammersley, Ian Hextall, Mary Hunt, Cathie Kennally, Dave Leon, Charlie Owen, Susan Powell, David Reynolds, Linda Santimano, Karen Scales, Stephen Shenfield, Tony Solomonides, Jane Steedman, Louise Stoll, Barbara Tizard, Monica Walker, Lorraine Wilder, Richard Winter, Michael Young.

Cover Design: Sue Ambrose
Inside Cartoons: Tineke Treffers

We invite readers to send comments and corrections in response to this pamphlet.

ISBN 0 09502541 9 3
Published by BSSRS Publications Ltd, 9 Poland Street, London W1V 3DC.
Typeset by Range Left Photosetters (TU) (01) 251 3959
Printed by Blackrose Press (TU) (01) 251 3043
Trade distribution by:
Southern Distribution, Albion Yard, 17 Balfe Street, London N1 (01) 837 1460
Scottish and Norther Books, Fourth Floor, 18 Granby Row, Manchester M1 3GE (061) 228 3003, and 48a Hamilton Place, Edinburgh EH3 5AX (031) 225 4950.

CONTENTS

PREFACE

This pamphlet is about educational research, and the direct and indirect ways it influences what happens in the classroom.

We have written it for an audience of 'consumers', some would say 'victims', of educational research results: teachers, administrators, parents, researchers, school governors, trainee teachers, lecturers in education and related disciplines – indeed, anyone who is interested in education.

Many people are bored, confused or downright dismissive of educational research, which seems to produce results that, to practitioners, are either obvious, or nonsense (or both), dressed up in incomprehensible jargon and obscure statistical techniques. The reaction is often just to ignore it as a specialised intellectual activity with no relevance to the everyday business of educating young people.

However such blithe dismissiveness can have dire consequences. Parents suddenly find that their own small local school which they know and approve of is to be closed and amalgamated with a larger school some distance away; teachers discover that their working practices are being checked up on more closely; or teachers' unions are told that class size makes no difference and that, if anything, pupils do better in large classes. All these are actions which have been supported by quoting the results of some particular pieces of educational research.

We wrote this pamphlet for those involved in education as we realise that educational research results are often used to silence the legitimate concerns of those wishing to speak up for their own interests. The use of 'statistics' and 'computers' is often thought to lend an aura of infallibility to research results so that those wishing to question the results are made to feel ashamed of their ignorance.

Perhaps a word about the authors is appropriate here. We are one of the subgroups of the Radical Statistics Group (for our address, see the inside cover), people who are interested in educational research, as practitioners and users. We have often been appalled at the way statistics have been used to justify conclusions and recommendations which will clearly harm the educational system. These feelings, and a commitment to helping practitioners build their competence and their confidence vis-a-vis reports of large-scale educational research, led us to write this pamphlet. We spell out our aims in detail in the Introduction.

One final word: though some of us are statisticians (by no means all!), we do not require our readers to be. Reading quantitative educational research may seem daunting to people who consider themselves to be 'without a maths background'. Our hope is that even the most technical issues raised here will be accessible to those who have not studied – or liked – maths very much in the past. If you know what means and correlation coefficients are, that is a help, but if not, there is a Glossary where all the technical terms used are defined. Also, we have listed the most readable statistics books we know in the Bibliography.

Radical Statistics Education Group
June 1982

INTRODUCTION

> 'The progressive primary school teacher, who lets children chatter, allows them choice of activities, never smacks them and rarely gives spelling tests, is due for an apology. kShe has spent five years standing disgraced in the class-room corner, after a major research project blamed her trendy, do-as-you-please methods for low standards in reading and maths. Now Neville Bennett, who carried out the research, says that it wasn't her fault after all and that her methods are just as good as the old-fashioned ones. It was, apparently, a case of mistaken identity'.
>
> P. Wilby, *Sunday Times*, 26 April 1981

Every year, it seems, newspaper headlines greet the results of the latest educational research project. The results are said, by those in authority, to be important, even definitive, although teachers and parents tend to be less impressed by them. Doubts soon follow – doubts expressed by 'experts' about methods, statistics and interpretation. Sometimes indeed, the original researcher pops up again to announce that he was wrong all along.

This cycle of events confuses many people who mutter 'why can't they make up their minds?' We hope, first of all, that this pamphlet will help readers make up their own minds. Second, we hope it will assist teachers and others to defend themselves against the use of research findings which purport to show that the failings of state education can be attributed to the performance of practitioners (and parents), while in fact these findings fail to take account of overall educational policies. These policies, both national and local, have resulted in services which, in many areas, are chronically under-financed. Thus, the broad aim of this pamphlet is to suggest, and portray, ways in which reports of quantitative educational research can be read more critically.

To try to achieve this aim, we develop ideas about the relationship of statistical methods to educational research and then we apply these ideas to three of the most influential and controversial pieces of educational research published in the last five years. These are Bennett's *Teaching Styles and Pupil Progress*, the *Fifteen Thousand Hours* study by Rutter and his colleagues on school effectiveness, and the National Children's Bureau study by Steedman, *Progress in Secondary Schools*. Sketches of these three studies are given in Chapter 1 and references to discussions of them are given in a Bibliography.

Chapter 2 starts with a general discussion about how we think statistical methods ought to be used in educational research. This covers both the production of research findings and the formulation of criticisms of these findings. We then go on to apply our views to the three case studies. In the course of our analysis, we adopt a very broad view of statistics and so include general questions of research methodology as well as the more traditional concerns such as statistical inference and model fitting. However, we do not attempt to provide a formal discussion of statistical techniques. Such discussions can be found in several text-books and the interested reader could refer to one of those listed in the Bibliography.

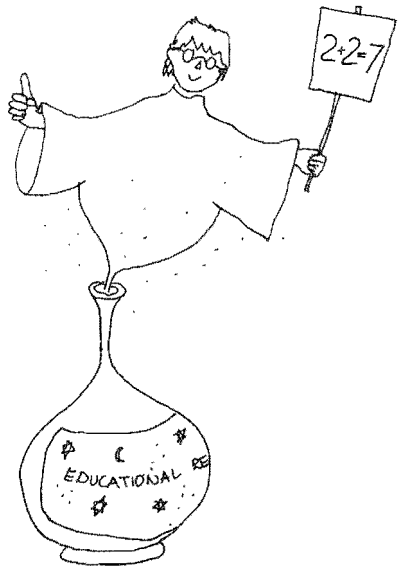
We recognize that educational research is more than just statistics and method, and that it is often difficult to distinguish statistical points from non-statistical ones. Our analysis of the relationship of statistical methods to educational research also considers other stages of the

research process. And so, in Chapter 3, we look at the formative stages of conceptualisation at the beginning of any project and the issues of presentation, interpretation and use at the end of the study. Again we develop our ideas in the context of our chosen examples, this time giving particular emphasis to *Fifteen Thousand Hours*.

We have concentrated on large-scale quantitative studies, partly because our own experience and competence is in this area and partly because these studies tend to impress and influence officials more than qualitative research. However, we do not wish to advocate the neglect of qualitative and ethnographic approaches particularly when these are used in conjunction with quantitative studies. (For pointers as to how this may be done, see e.g. Evans (1982b), and the references therein.) Of course, qualitative research must be examined just as critically as quantitative research although the tools for such an examination will inevitably be different and a discussion of them is beyond the scope of this pamphlet.

Another area which we do not cover is that of official British statistics of education. Readers interested in a critical appraisal of these are referred to Fielding (1981) while critiques of official statistics of all kinds can be found in Section 3 of *Demystifying Social Statistics* (Irvine, Miles and Evans, 1979).

Educational research tends to produce educational myths, the 'failure' of informal teaching styles in primary schools being one of the more powerful in recent years. Our hope is that readers can use the guidelines given in Chapters 2 and 3, when faced with having to evaluate future studies of this kind. We would like to see a more critical audience for educational research which relies heavily on quantitative techniques. Such an audience might discourage myth-making and encourage fair and sensible educational policies based on sound research.



—EDUCATIONAL RESEARCH TENDS TO PRODUCE EDUCATIONAL MYTHS —

1 OUR THREE CASE STUDIES

1.1 INTRODUCTION

All three of our case studies have characteristics in common. They all employ relatively sophisticated quantitative and statistical techniques and each of them was subject to varying degrees of criticism about the use of these techniques. They are all based on essentially the same research design which compares groups – formal, informal and mixed teaching styles (Bennett), twelve secondary schools (Rutter et al.), and selective and non-selective secondary schools (National Children's Bureau). The formation of these groups was not under the control of the researcher (as it would be in an experiment) and members of the groups were followed up over time. In other words, they used a quasi-experimental, longitudinal design (see Glossary). Also, they are all related to issues of educational policy.

1.2 CASE STUDY NO.1: *Teaching Styles and Pupil Progress*

The 1960s were a decade of great change in English primary education. Official policy aimed to make primary schools more human places; teacher training courses emphasised informal methods of teaching; and the Report of the Plowden Committee (1967) put a strong case for 'discovery' methods which, it argued, would prepare children for a society where knowledge and skills would be constantly changing.

These Plowden-related perspectives were a radical departure from tradition. Parents, brought up in a society where facts were facts and had to be learned, and where no medicine could do you good unless it was unpleasant, were naturally worried when their children seemed to spend an inordinate amount of their time at school simply 'playing'. With hindsight, it is clear that many schools should have spent more time explaining to parents the value of such radical departures, and in particular the apparent value of *planned* play as an effective learning method.

It was in these circumstances that the Department of Educational Research at Lancaster University obtained a grant from the Social Science Research Council (SSRC) to study, among other things, the effectiveness of different styles of teaching. The project was co-directed by Neville Bennett, now Professor of Educational Research at Lancaster. After an extensive review of the research on teaching styles, he designed a questionnaire which, he hoped, would identify important aspects of 'progressive' and 'traditional' styles. This questionnaire was completed by 468 third year junior school teachers in Cumbria. Their answers were then analysed to determine 'natural' groupings or clusters of teachers from the data, each cluster corresponding to a different teaching style. This technique is known as cluster analysis (see Glossary).

The particular clustering method used grouped the 468 teachers into 12 clusters. Of these, two clusters were classified as being 'formal', two as 'informal' and three as 'mixed' (the others remaining undefined). A total of 37 teachers were selected as being representative of the three types of teaching style. All the pupils in the classes of these teachers were tested

in mathematics, reading and English at the beginning and at the end of their fourth primary year.

The research team then carried out an analysis, to try to assess the influence of teaching style on the progress of the pupils over the year. The type of analysis they used is often referred to as an analysis of covariance. They found that, overall, the formal classes made the most progress, informal least, with 'mixed' classes somewhere in between and they claimed that these results formed a coherent pattern and were statistically and educationally significant. The book – *Teaching Styles and Pupil Progress* – was released in a blaze of carefully orchestrated publicity, even to the extent of having an entire BBC *Horizon* programme devoted to the results on the day of publication. The anti-progressive lobby, as might have been expected, seized upon these results to justify their campaign to such effect that at least one educational journalist (Peter Wilby, the education correspondent of the Sunday Times, in the article from which the quotation in the Introduction is taken) identified the book as a major factor in the anti-progressive backlash of the 1970s. Indeed Bennett's findings made such an impact that James Callaghan, the then Prime Minister, referred to Bennett's conclusions, (while not actually mentioning him by name) in his speech kicking off the 'Great Debate' on education conducted in 1977/78.

1.3 CASE STUDY NO. 2: *Progress in Secondary Schools*

Since the start of the national reorganisation of maintained secondary schools in the 1960s, there have been attempts to compare the performance of comprehensive, grammar and secondary modern schools. Most of these have involved comparisons of pupils' academic performance, and typically have used public examination results. The difficulties with this approach may be summarised quite simply. Because the grammar schools select the most academically able children in an area, one would expect their pupils to have, on average, a better academic performance by the end of their schooling than those who go to unselected comprehensives, even if the quality of schooling in the two types of school is exactly the same. Attempts have been made to allow for this so-called 'creaming' effect but none have been successful. Further discussions of some of the technical problems can be found in Baldwin (1977) and Goldstein (1977).

In 1977, the Department of Education and Science (DES) funded a project at the National Children's Bureau, using data from the National Child Development Study, which aimed to compare pupils' performance in different types of school, statistically adjusted to take into account their measured performance prior to starting secondary school. The report of the study is an extensive document which looks at factors associated with a number of 'outcome' variables. We shall concentrate on just two of the outcomes, mathematics and reading attainment, since these have aroused the most public interest. However, the report also presents results for outcomes such as attendance, plans and aspirations.

The study identified a group of schools which had been either entirely comprehensive, entirely grammar or entirely secondary modern during the time the children in the study cohort attended them. These children belonged to the birth cohort born in 3-9 March 1958, so that the period of study, 1969 to 1974, was one when some comprehensives had been well established whereas others were new, but there were also many selective schools still in existence. This means that the results are strictly relevant to that historical period only and

do not necessarily generalise to other situations. Nevertheless, it seems clear that these data are the best which anyone has been able to analyse up till now and the results were therefore almost certain to become important politically. The main findings are as follows.

At the age of 11, before transfer to secondary schools, children completed tests of reading comprehension and mathematics. Not surprisingly, those going on to grammar schools had higher mean scores on both tests than those going on to comprehensives who in turn had higher mean scores than those going on to secondary moderns. At the age of 16, in their last year of compulsory schooling, they completed the same reading test and a different test of mathematics. The same pattern emerged. However, one could say that the differences at 16 resulted from differences at 11 rather than from the type of school the pupil attended. Therefore, the mean 16 year scores for the three types of schools were compared for children who had the same 11 year score and were similar on other variables too. This involved examining 16 year scores for each separate 11 year score within selected subgroups of the population. In other words, triplets of children were constructed in the analysis so that each child was alike in as many respects as possible but one child from the triplet went to a grammar school, one to a comprehensive school and the other to a secondary modern school. (In statistical terms, an analysis of covariance or multiple regression was carried out on 16 year scores holding 11 year score and other variables constant.)

An interesting pattern emerged from the comparisons. For those children with *low* 11 year scores (who did not go to grammar schools), there was generally little difference in 16 year scores between the comprehensives and secondary moderns. For those children in the *middle* score range at 11 this was also true, but those who went to grammar schools did better than the other two groups at 16, given the same 11 year score. Thus one may conclude that for children in this middle group being in a grammar school resulted, on average, in greater academic progress. For those with the *top* 20% of marks at 11, the progress of those in grammars and comprehensives was very similar and higher than that in secondary moderns. So, for these most able children, similar progress was made in comprehensives and grammars and more than in secondary moderns.

1.4 CASE STUDY NO.3: *Fifteen Thousand Hours*

Parents who are concerned about their children's progress, will answer 'yes' unhesitatingly to the question 'Are some schools better than others?'. Indeed, they will often make great efforts to ensure that their children go to the best available school, even to the extent of moving house to ensure that they are in the right catchment area. Teachers too will answer 'yes', though some will be aware that not everyone seems to agree. Certainly influential American research such as Coleman et al. (1966) and Jencks et al. (1973) seemed to indicate that schools account for only a small part of the variation in pupil attainments. And, on this side of the Atlantic, the Plowden Report, in some ways an English counterpart to Coleman's work, came to the conclusion that home influences far outweigh those of the school.

It was in this context that the Inner London Education Authority (ILEA) and DES funded Michael Rutter and his colleagues at the Institute of Psychiatry to investigate school influences on children's behaviour and attainments, paying particular attention to those aspects of schools which influence their functioning as social organisations. The study looked at 12 non-selective ILEA secondary schools and the progress of a year-group of

pupils within each school. One of the great strengths of the research was that it was longitudinal and thus able to control for differences in the intake of the 12 schools in a way which previous studies had been unable to do.

The study's outcome variables were school attendance, attainment in public examinations, school behaviour and delinquency. Inferences about these variables were made after attempting to allow for differences between the intakes of the schools in terms of verbal reasoning test scores at age 10, parental occupation, scores on a behavioural questionnaire completed about their pupils by primary school teachers and whether the children's parents were immigrants to Britain. The study team also collected information about the physical and administrative characteristics of the schools such as number of pupils, pupil-teacher ratio, number of sites and sex composition. However, the major way in which the study was innovative was in its attempts to document the style of social interaction within the schools; this information on school process was collected by a combination of interviews with staff, a questionnaire to pupils and systematic observation of a complete week in each school with middle ability third year classes.

The main findings of the study were as follows:-

(i) The schools showed marked differences on each of the four outcome variables and these differences could not be explained by differences in their intake: pupils' experiences at secondary school may influence their progress.

(ii) Differences between schools were systematically and, probably, causally related to their characteristics as social institutions – their 'ethos' – but not to their physical and administrative characteristics.

(iii) Outcomes were also influenced by extraneous factors, particularly the academic balance of the intake.

The publication of the book from the study generated a good deal of interest and publicity: rather more than the National Children's Bureau study but perhaps less than Bennett's book. The finding that schools can and do make a difference was a popular one. In addition, the argument that these differences seemed likely to be caused by factors within the school was seen to be helpful in that it suggested ways in which teacher, heads and administrators could get the maximum impact from their work. However, more considered assessments of the findings led to the expression of doubts and criticisms which we come back to later.

1.5 SUMMARY

In this Chapter we have described three case studies which claim to address major issues of educational policy: formal vs. informal teaching methods (Bennett), selective vs. comprehensive secondary schooling (National Children's Bureau) and school effectiveness (Rutter et al.). All three studies received substantial publicity and have generated controversy both in the academic community (see Section 2 of the Bibliography), amongst those concerned with education, and in the wider society. Their conclusions are still being cited and critically discussed, and their implications touch some of the most cherished beliefs and practices of various sectors of society.

We now go on to look critically at the methods and findings of each.

2 USE AND ABUSE OF STATISTICAL METHODS

2.1 INTRODUCTION

In this chapter, we look at some of the ways in which statistics and educational research interact. Our perspective does not suppose that there is always just one correct statistical approach to a research question. Indeed we recognize that statistical analyses are based on assumptions and choices which are often arbitrary. Nevertheless, it does seem to us that statistical methods are often applied improperly to educational data and this is also true of many statistical criticisms of educational research (including some we have made ourselves in the past!). Therefore, we put forward a series of criteria, in the form of questions which we think should be asked of research and which can be seen as defining, in a rather loose way, the idea of 'statistical responsibility'. These questions are essentially technical ones and we separate these from other, non-statistical, questions which can be asked of research although we shall have occasion to query the neatness of this distinction in the next chapter.

2.2 TOWARDS STATISTICAL RESPONSIBILITY

Before listing these questions, let us look in a little more detail at the function of statistical criticism. Our thinking on this matter began when we were attempting to distinguish worthwhile criticism of *Fifteen Thousand Hours* from criticism which seemed rather unfair. We found ourselves in agreement with Bross' (1960) claim that critics should aim to clarify rather than to obscure the issues underlying the research in question.

Bross puts forwards the idea of a 'hit and run critic' who 'points out some real or fancied flaw and supposes that his job is done' despite the fact that the overall effect of the error could not invalidate the conclusions. For example, though it would be careless to publish a standard deviation in which a mistake had crept into the calculation, it would be irresponsible to condemn a piece of research on the basis of one wrongly calculated standard deviation.

To avoid the proliferation of hit and run criticism, Bross proposes that a statistical critic should present a counter-hypothesis to the one being advanced by the proponent. The counter-hypothesis should be 'tenable' which means that it should satisfy the following double requirement:-

(i) it should be plausible theoretically; and

(ii) it should be confirmed empirically by the data under discussion or by some other relevant data.

To satisfy (i), critics must make explicit the concepts and theories they are working with[1]. And, of course, requirement (ii) can only be met if the proponent's data are publicly available.

We also drew on the work of Cook and Campbell (1979), following Campbell and Stanley (1966), who argue that it is the proponent's job to recognize plausible counter-hypotheses in advance, and to attempt to control for these in their research designs in order to make their

conclusions valid. Cook and Campbell refer to four kinds of validity – statistical conclusion validity, internal validity, external validity and construct validity, which we shall explain below.

A combination of these two sets of ideas led us to a list of questions and guidelines which could be addressed by those concerned with evaluating research. Readers may find the list incomplete and we would encourage you to add your own. We also think these questions could usefully be considered by those engaged in educational research, particularly when they are writing up their results (see Evans, 1982a).

1. Are the measures or indicators chosen to represent the underlying concepts of the research appropriate? In other words, do they have *construct validity?*
2. Have the researchers taken into account the effects of sampling error on their conclusions? In other words, is it likely that they would have come to the same conclusions if, by chance, they had selected a different sample? If critics can demonstrate that the results can in fact be explained by chance then they are, in our view, presenting a tenable counter-hypothesis. This is the issue of *statistical conclusion validity.*

   The most common method of dealing with sampling error is through the use of tests of statistical significance. Have the researchers used these properly? In particular, have they understood the distinction between statistical significance and educational significance? In studies with large samples, differences and associations can be highly significant statistically but their magnitudes can be small and educationally uninteresting. Conversely, in studies with small samples, tests are likely to fail to reject the null hypothesis of, say, no difference when there really is an important difference in the population. In other words, the tests have low 'power' when samples are small.

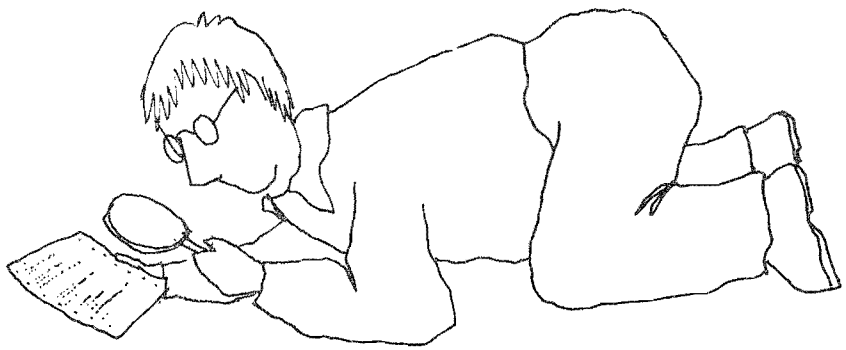
   We would not regard as reasonable, criticisms which merely pointed out that tests of statistical significance had not been used or had been used wrongly. The critic must show that different conclusions would have been reached if the right tests had been used.[2]
3. Are there alternative explanations involving 'third variables' that could invalidate the findings and that have not been taken into account or 'controlled' (see Glossary)? This is the question of *internal validity.*

   We stress the need to focus on real threats to validity, in other words, *tenable* counter-hypotheses. For example, an association or correlation between two variables is certainly not a sufficient condition for there to be a causal relationship between them. On the other hand, little is gained by critics who say merely that association is not the same thing as causation; they need to say which other variables might have been included in the underlying statistical model and how these changes might have affected the results. This is particularly important when discussing the threat of 'selection' when groups of people receiving different 'treatments' have different characteristics (as in quasi-experiments).
4. Is the sample properly described and are unjustified generalisations made about populations which were not sampled? This is the question of *external validity.*
5. Is the analysis conducted at the right *level*? Is it pupils, classes or schools or some other level that is of interest? Would the results be affected by an analysis at a different level or by an analysis which explicitly took account of more than one level? This is particularly important in educational research with its objects of study hierarchically ordered from pupils to classes to schools and on to local education authorities (LEAs).

6. Are the statistical procedures *properly explained* so that non-statisticians can come to their own conclusions, however tentative, about whether the chosen methods were in fact appropriate to answer the research question posed?

   It is particularly important that the advantages and disadvantages of less familiar and more complex techniques are explained. We recognise, however, that publishers and journal editors may need to be persuaded of the importance of such material.
7. Are there *discrepancies* between careful interpretations and qualifications in chapters giving results, and wide claims in the conclusions?
8. Is there any firm evidence of *fabrication* of data or *deception* in the way the results are presented? If so, this must be exposed whatever the consequences for the research.

We now turn to our three case studies. First, we look at how statistical methods were used in Bennett's research and how they have been shown to have been unsatisfactory; this raises the question of statistical responsibility in the case of the proponent of particular findings. Second, we consider certain criticisms made of the research on selective and non-selective schooling in the light of our criteria developed here. Finally, we criticise *Fifteen Thousand Hours* in a way which we consider to be responsible and constructive.



– IS THE ANALYSIS CONDUCTED AT THE RIGHT LEVEL –

2.3 CASE STUDY NO.1: *Teaching Styles and Pupil Progress*

The results of Bennett's study, that pupils make most progress under formal teaching styles, and especially the certainty of tone with which they were presented, conflicted with the experience of many practitioners and with the beliefs of many people involved in education and it was thus inevitable that the study should be subject to close critical examination. Critics directed their attention to two of the statistical aspects of the study. First they argued that observed differences in progress between the three teaching styles, which Bennett claimed were causal, could equally well be attributed to other characteristics of the teachers, their classes or their schools. This relates to question 3 above, concerning internal validity. Second, they argued that the results really depended on just 37 classes and teachers rather than each of the 950 children so that the differences, far from being reliable, were in fact based on a small sample and could be explained by chance. This relates to questions 2 and 5 in the previous section concerning statistical conclusion validity, and the appropriate 'level' of analysis. It is interesting to note that few criticisms were made of the way in which teachers were grouped, using cluster analysis, into the various styles and one might speculate that the apparent sophistication of this technique deterred the critics, few of whom were statisticians, from attacking it. We do not give a detailed discussion of these criticisms here because the data were in fact re-analysed to try to deal with some of them; instead we focus on a comparison of the two analyses.

The controversy and criticism generated by the original study eventually led Bennett and Professor Murray Aitkin of the Centre for Applied Statistics at Lancaster University, to apply to the SSRC for a grant to re-analyse the data. This project, under Aitkin's direction, concentrated on the development of more appropriate statistical methods for the analysis of the original data.

The results of the re-analysis differed from those originally reported by Bennett. They are reported in a relatively non-technical form in Aitkin et al. (1981), and discussed by Gray and Satterly (1981) in the same issue of the journal. They may be summarised as follows: *assuming a clustering model was appropriate*, Aitkin provides evidence for the existence of 3 groups of teachers, but these groups differ somewhat from Bennett's 'formal', 'informal' and 'mixed' groups. Aitkin's first goup was closely similar to Bennett's group, and both can be designated as 'formal'. The designation of Aitkin's second group as 'informal' seems to have construct validity when we look at responses of these teachers to the various questions designed to measure teaching style – but its membership is not the same as Bennett's 'informal' group. The third group produced in the re-analysis differed considerably from Bennett's third group who had been described as 'mixed', but had seemed in some ways to combine many of the advantages of 'formal' and 'informal' styles. Aitkin's third group, on the contrary, comprises those teachers who were *least* successful in using a consistent style in the classroom; they had more discipline problems, and seemed less systematic in their approach as they were lowest in giving homework or regular tests or assessments.

The second part of the re-analysis was concerned with the progress made by pupils in the classes of the three groups of teachers. The results differ from Bennett's in two ways: first, the patterns of differences between the groups on the three outcome measures is different and second, none of these differences are statistically significant. This last finding is probably the most important way in which the findings of the re-analysis differ from the original research. It occurred because the technique Bennett used made the unreasonable assumption that none of the formal teachers say, was any more or less effective than any of the other formal teachers. This means that he assumed that, within a teaching style, there were no differences in the way the children were taught at school. In other words, Bennett ignored the variation among teachers apart from the variation attributable to the three categories of formal, informal and mixed. Describing these findings in another way, one could say that even had Bennett got his teacher clustering 'right', the differences between styles would have been too small to be statistically significant, and so chance would indeed have been a tenable counter-hypothesis for the explanation of differential progress.

Everyone interested in the effect of different teaching styles is now faced with two interpretations of the Bennett data. What are they to make of them? They may know that Bennett's analysis was heavily criticised whereas Aitkin's re-analysis has attracted little critical comment and none in the educational press and journals. However, we do not think this is a good criterion for choosing between the two analyses, partly because the re-analysis involved the construction of rather complicated statistical models and potential critics might not have had the confidence to criticise them. Let us consider each of the three important statistical questions separately.

First, was Aitkin's clustering approach better than Bennett's and, if so, why? Or should we still question whether any kind of clustering approach is sensible? It was well-known that the clustering methods available at the time Bennett was analysing his data were not to be relied upon. Both the number and the membership of clusters could vary according to the numerical methods adopted. However, readers of *Teaching Styles and Pupil Progress* were given no intimation that the cluster analysis was anything but a tried and tested technique and we believe that this was not responsible. Aitkin's method is based on statistical theory in that teachers were allocated to clusters for which they had the highest probability of belonging. Nevertheless, as Aitkin et al. point out, 'differences remain in deciding on the number of clusters'. It is much more realistic to suppose that teachers vary continuously on the formal/informal dimension and that teaching style is in fact a multi-dimensional concept. However, further analysis of the Bennett data set may never be able to answer these questions. Its narrow focus on formality and informality means that we cannot carry out the essential process of testing and comparing other possible explanations since the information to do so is lacking, and since we can never go back to the mid-1970s to ask teachers further questions.

We believe there are two lessons for the general reader from this:-

(i) be suspicious of unfamiliar techniques which are not explained (question 6).

(ii) always look back at the original questions asked, to see how well they capture the underlying variables of interest. This is very much the issue of construct validity (question 1).

Second, we believe that Aitkin's inclusion of a variable which allows for variation between teachers within teaching styles is a major improvement on the original analysis. The lesson here is that analyses of educational data which ignore the structure of the situation being studied are likely to mislead, and this is the point of question 5. The issue is not just at what level to analyse the data (for example, child or teacher) but also the inclusion of all potentially important levels in the analyses.

Third, the issue of uncontrolled variables explaining the differences in progress between styles is left unresolved by the re-analysis. Perhaps this no longer matters because these differences can now reasonably be explained by chance. However, those readers who

believe the magnitudes of the observed effects are still educationally interesting must remember that there may be variables associated with style and with progress over the year after allowing for the pretest, which could account for these differences. In other words, the findings that the observed differences are due to teaching styles may not be internally valid (question 3). A list of plausible omitted variables would include teacher age and experience and the effects of the 11+ on style and curriculum and thus on progress in reading, maths and English.

2.4 CASE STUDY NO.2: *Progress in Secondary Schools*

Progress in Secondary Schools claimed that there were few differences in pupil progress between selective school systems and comprehensives. Two months after the report came out, a long critique was published (Cox and Marks, 1980). In this, the authors took up a number of points and we shall examine three of these. First, they criticised the report for failing to set out the 'raw data' on which the findings were based and they subsequently elaborated on this in press correspondence. Second, they criticised the reading test used, in particular on the grounds that it had too much of a 'ceiling' at 16 years, i.e. too many children were obtaining full or nearly full marks to make it a sensitive discriminator between types of school. This is an issue of construct validity. Third, they accused the author and research team of emphasising only those results which conformed to their predetermined view and thus presenting a biased and distorted report, summary and press release. This relates to questions 7 and 8 of Section 2.2. We shall consider these three criticisms in turn and use them in order to make some general remarks about responsible criticism.

(i) When talking of access to 'raw data', Cox and Marks seem to refer sometimes to access to *individual* data on magnetic tape, and sometimes to the presentation in the report of *unadjusted* frequency distributions of 11 year and 16 year scores. So far as access to the magnetic tape is concerned, the National Children's Bureau did in fact follow common practice in depositing the tape with the SSRC Data Archive. It was made clear in letters from researchers (*Times Education Supplement* [*TES*], 10 Oct. and 17 Oct. 1980) that the Archive did not refuse access to its data for bona fide researchers. In this connection, it is interesting to note that Cox and Marks did not seek access to the data after this misunderstanding was cleared up.

On the presentation of detailed frequency distributions and other basic data, in published reports, there is an interesting area of debate. On the one hand, for experienced researchers such extra information about, for example, unadjusted scores at 16 might enable them to gain further insight into the results. On the other hand, publishing the unadjusted results might have focussed attention on the unsurprising finding that grammar school pupils score higher than comprehensive pupils at the end of their secondary schooling. But this is quite irrelevant to the study's own focus on the *progress* that children made during their secondary schooling. A more reasonable criticism in this context might have been that full distributions of 16 year scores were not published for given 11 year scores but this was not the criticism made.

Thus, although some disagreement here is possible, any dispute is essentially marginal. Unless the critics argue that the omission of such data actually threatens the internal validity of the conclusions (and Cox and Marks do not claim that), the criticism is more in the nature of a debating point and in our view irresponsible. They made considerable use of this criticism in press correspondence, leaving the impression that they considered it a potential threat to validity, whereas in fact it amounted to nothing of the kind. We return to the issue of presentation in Chapter 3.

(ii) Cox and Marks severely criticised the reading test used at both 11 and 16 years on the general grounds that no such test could be appropriate to both ages, and in particular that at 16 years there was a bunching of scores near the top of the scale. Thus for the high ability 11 year olds in particular, the test would be very poor at discriminating between different types of school, so that the resulting lack of difference between comprehensives and grammars was unsurprising.

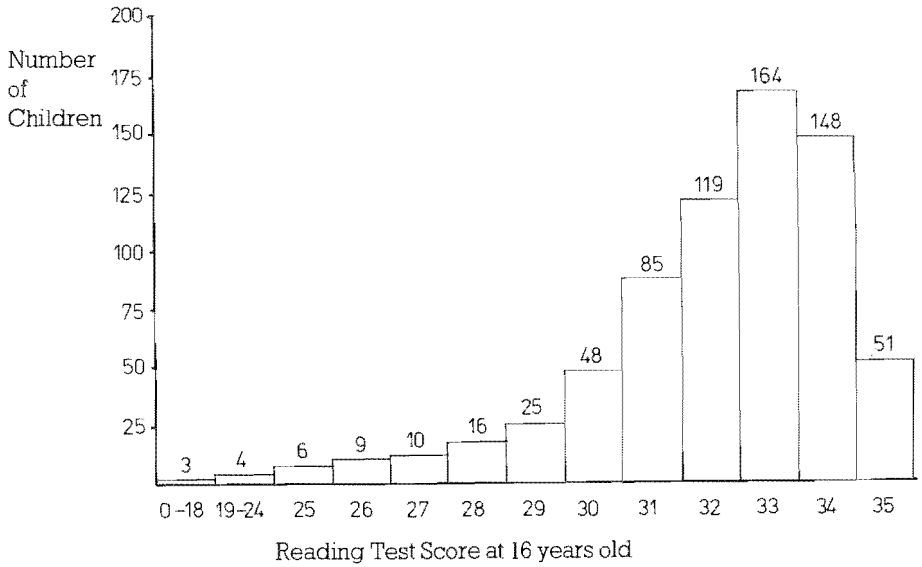
This criticism is more substantial than the previous one since it does, prima facie, present a threat to validity. Indeed, Steedman mentions this difficulty with the test and she argues that it does not in fact invalidate the results. No research in education can be completely without areas which are open to discussion, but the research can only identify these and present arguments about which assumptions are likely to be reliable and to what extent. While it is important to identify such problem areas and to assess the validity of the research arguments a critic must also have a sense of their overall importance in the study. Cox and Marks in describing the use of the reading test at 16 say: 'The National Children's Bureau knew the test was virtually useless at 16 for high attainers at 11 but presented the results in a way which concealed this fact.' In fact, this was not so. Nevertheless, in their published rebuttal of Cox and Marks' allegations, Steedman et al. (1980) do show that the top-scoring eleven-year olds had high scores at 16 although very few of them gained the maximum score on the test (see Figure 2.1). Thus, there must be some doubts about the validity of the reading test and the report's view that they are not sufficiently serious to affect the findings is a judgement and could be wrong. On the other hand, Cox and Marks' criticism of the mathematics test at 16 is unwarranted as Steedman et al. demonstrate (see Figure 2.2).

(iii) According to Cox and Marks the report 'is so biased in its interpretation of its own data that it is hard to avoid the suspicion that those concerned with its production, including the Advisory Group on which the DES were represented, were capable of gross partiality and/or influenced by vested interests'. In subsequent letters to the Press (e.g. *TES*, 3.10.80) this point was reiterated.

This raises an important point about the meaning of the word 'bias'. All social research gives rise to results which may be open to differing interpretations. Different researchers confronting the same results might well give different emphases or come to different conclusions which will depend on their own experience and theoretical position. To call this 'bias' with the popular overtones which that word has, is misleading since it is to hanker after 'value-free' research which is impossible (see Chapter 3).

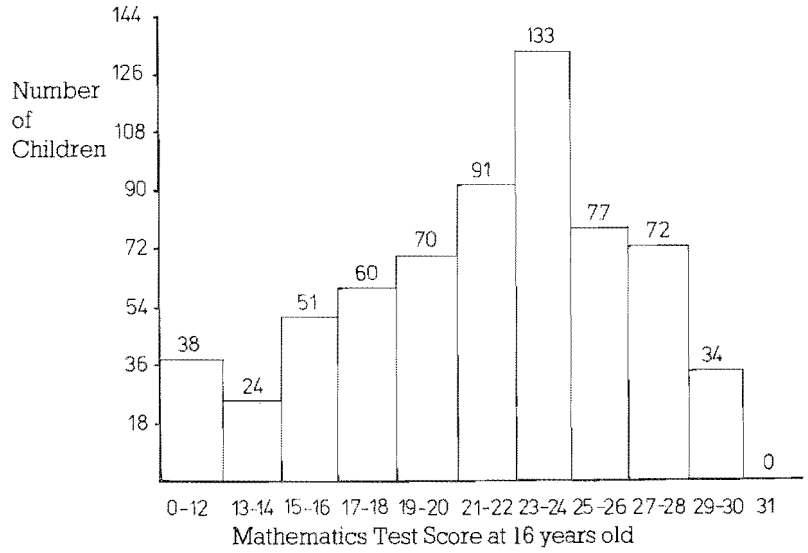
It is interesting to note that a statement by the British Educational Research Association ('Research Intelligence', April 1981) criticised Cox and Marks for 'language and tone which went beyond what is normally acceptable'. Rather than impugning the researchers' motives, Cox and Marks could have made a genuine attempt to re-analyse the data in order to provide a tenable counter-hypothesis rather than just irresponsible accusations. A real source of bias and vested interest can be seen in a report which includes sentences such as '(the report) implicitly advocates the extension of comprehensive schooling beyond the *already frighteningly high* figure of 85% of our nation's children' (our italics), clearly indicating where they stand with or without research.

Figure 2.1 Reading Scores at 16 years for Top-Scoring 11 year olds



Source adapted from Steedman et al. (1980), p.2

Figure 2.2 Mathematics Scores at 16 years for Top Scoring 11 year olds



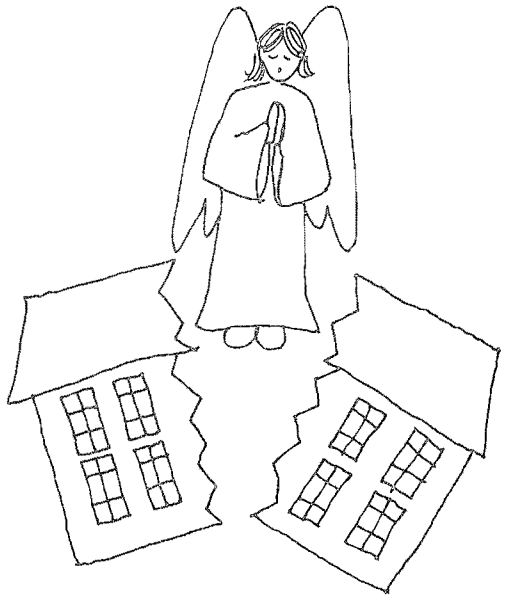
Source adapted from Steedman et al. (1980), p.3

2.5 CASE STUDY NO.3: *Fifteen Thousand Hours*

Fifteen Thousand Hours suggested that school processes are important to the progress, attendance and behaviour of children. It is certainly an interesting and innovative report which resulted from a considerable amount of careful and original work and which broke new ground in the study of the determinants of educational success. However, serious reservations about the study have been published in reviews and in two pamphlets, one by Tizard et al. (1979) from the University of London Institute of Education and one by Wragg et al. (1980) from Exeter University. We have drawn on the work of some of these critics in formulating our own statistical criticisms. We do this by working through each of the eight questions given earlier in the chapter.

1. There must be serious doubts about the construct validity of the ethos variable. Ethos was measured by 39 of 46 indicators for school process, the 39 being selected on the basis of a statistically significant relationship with one of the dependent variables. Not only is this mechanical approach unlikely to be a valid way of specifying indicators for the ethos concept, the fact that this created a variable which is highly correlated (0.92) with school behaviour, for example, provides no basis for inferring the *existence* of an ethos. The fact that the concept of ethos is so nebulous (see the chapter by Dancy in Wragg et al.) is another reason to argue that the construct validity of the indicators chosen to measure it is low.
2. The book relies heavily on the results of significance tests even though the samples studied could in no sense be regarded as random samples from well defined populations which the textbooks insist is necessary. Nevertheless, on the whole, one could not argue that the findings on school effects are merely due to chance.

   However, the authors dismiss potentially interesting findings about the physical and administrative features of the schools and about ecological influences just because they are not statistically significant at the arbitrary 0.05 level. As we pointed out in Section 2.2, significance tests on small samples have low power and all the tests reported in *Fifteen Thousand Hours* for looking at school-based variables are calculated on samples of 12 and sometimes even fewer. This means that they would tend not to reject hypotheses of no differences or no association even when these null hypotheses were false (i.e. when there was an association in the population). For example, when the twelve schools were ordered from lowest rates to highest rates for attendance, the mean rank of the four voluntary aided (i.e. Church) schools was 4.0 and the mean rank of the eight local authority schools was 7.8: this is rather a large difference and although not significant at the 0.05 level, it is significant at the 0.08 level and there is a good case for raising the significance level when the samples are as small as they are here. Similarly, for those ten schools for which there was information on parental choice, the mean ranks for attendance were 8.7 for relatively unpopular schools, 5.0 for schools with average popularity with parents and 3.5 for popular schools and these differences were significant at the 0.10 level although again not if 0.05 is used as the criterion. When these results are considered along with the analyses of the effect of split sites which showed that schools on split sites had *significantly* better behaviour and *significantly* less delinquency then we are able to put forward a tenable counter-hypothesis: the differences between the schools could be explained at least in part by some of their physical and administrative features and by parental choice.



— SCHOOLS ON SPLIT SITES HAD BETTER BEHAVIOUR —

3. The authors properly recognized that observed differences in outcomes between the schools would only be credible as school effects if initial differences between the intakes of the schools had been eliminated in the analysis. However we argue below that the ways in which allowance was made for differing intakes was not satisfactory although we are not in a position to put forward a really tenable counter-hypothesis partly because the authors have not yet made the data available for re-analysis.

   Readers of *Fifteen Thousand Hours* are likely to be confused, as we were, by the way in which the authors control for verbal reasoning at intake; sometimes they use a variable divided into 7 groups which is acceptable but sometimes they use just a 3-band variable with half the pupils in the middle band. This is much less satisfactory because schools could differ considerably in verbal reasoning intake and still have equal proportions in the middle band. The authors do control, although this is not explained with total clarity, for two home background factors – social class and country of origin. However, it is possible that the inclusion of other factors such as parental expectations and interest and family size, which were not available to the research team, would have eliminated the observed school differences.

   Tizard, in the Institute of Education publication (p.26), makes the point that 'some working-class parents with children of average ability are more knowledgeable about and interested in education than others. If these families select a secondary school with a "good" reputation, and thereafter give their children more educational support, the children's school career will depend to a greater extent than the authors allow on parental as well as school characteristics'. In addition, Hutchison et al. (1979) show that family size and housing have independent effects, although admittedly small ones, on

   achievement at 16 after controlling for achievement at 11 and social class. We are inclined to believe that school differences in the four outcomes would still have been found with more adequate control for intake but we are sure that the magnitude of these differences would have been reduced. (See Gray [1981] for essentially the same argument.)

4. There is a clear description of the schools sample and the authors point out that schools in the ILEA area were, at the time, relatively well-off in terms of resources although they did suffer from high staff turnover and were dealing with reorganisation, industrial disputes and the prospect of falling rolls. This means that any generalisation to schools in England or in Britain can only be very tentative, yet the authors sometimes fail to acknowledge this when considering the implications of and conclusions from their study. The authors do provide evidence from other studies to support their claims but the external validity of *Fifteen Thousand Hours* must be considered rather low. It needs to be extended by further studies in other parts of Britain.
5. As noted in 2., some analyses were done on the sample of children (n=2200) and some on the sample of schools (n=12). The authors did not allow for variation between teachers within schools but this would have been difficult and we would not criticise the research on the question of appropriate levels of analysis.
6. We need to consider whether the relatively unfamiliar statistical techniques which were used in the analysis are explained in a way which non-statisticians could understand, at least at an intuitive level. In particular, the method of log-linear modelling is applied to some of the data; this is a valuable technique which has become popular with statisticians over the last decade but which is not yet familiar to non-statisticians. Appendix H in *Fifteen Thousand Hours* tries to explain the technique to a lay audience but it employs a lot of rather complicated algebra which will be difficult for most readers to grasp. We recognize the difficulties of explaining statistical techniques to the general audience but statisticians and others who use the techniques must try to put them across in as clear and as simple a way as possible so that readers can begin to evaluate their strengths and weaknesses rather than just accepting them because 'experts' have chosen to use them.
7. Did the authors use their concluding chapter to make sweeping statements which are not justified by the data? We have already pointed out that the external validity of the study is low but that this is not reflected in the conclusions. The authors are careful to point out the tentative nature of their results in most of the book and they do say in the final chapter that 'firm conclusions about causation can only come from controlled experimental studies.' However, the rest of the chapter does assume that the correlations can be translated into causal statements. It is reasonable that studies like this (and like Bennett's and the National Children's Bureau's) should try to produce more than just descriptive statements but caveats about causation should appear with the conclusions. After all, many readers will concentrate on the first and last chapters and will thus miss the fine detail of the analyses.
8. There is no evidence of fabrication or deception in the way the results are presented.

To conclude, we believe there are doubts about the magnitude of the school effects found in *Fifteen Thousand Hours*, and that the physical and administrative variables do provide a plausible counter-hypothesis to school processes for the explanation of the school effects. In any case, the external validity of the study is low.

2.6 SUMMARY

This chapter presents a notion of *statistical responsibility*, to help critics of research reports avoid what has been called 'hit and run', or destructive, criticism. This approach recommends the critic to present criticisms as counter-hypotheses which should be both plausible theoretically, and supported by the data under consideration or by other relevant data.

We present a list of questions for readers to consider when critically reading a research report. These questions have to do basically with the statistical and methodological qualities of the research, and questions 1. to 5. can be thought of as areas which might generate tenable alternative hypotheses. The eight questions are concerned with:

1. construct validity;
2. statistical conclusion validity;
3. internal validity and 'third variables';
4. external validity;
5. levels of analysis;
6. explanations of statistical techniques appropriate for the intended audience;
7. discrepancies in the care given to interpretation in different parts of the report; and
8. fabrication of results and deception in presentation.

To stay with these questions as the only basis for criticism, tends to ignore other issues. For this reason, in the next chapter we develop a wider perspective from which to read research critically.

End-Notes for Chapter 2

1. We would expect proponents, too, to make their concepts and theories explicit and acknowledged.

2. In the same way, critics who argue that the chosen method of analysis is inappropriate, for example that log linear models rather than multiple regression should have been used, need to present convincing reasons for supposing that different results would have been obtained with the 'right' technique.

3 BEYOND USE AND ABUSE OF STATISTICAL METHODS

3.1 Introduction

The intention of the ideas presented in the previous chapter is to specify the appropriate scope of statistical and methodological evaluation. That is, basically, the methodologies and techniques used in the design, execution and analysis stages of the research. However, this focus overlooks both the formative stages and the stages of application of the research. The formative stages include the selection and conceptualisation of the research problem, and the allocation of resources to it. A perspective which concentrates only on statistical aspects will tend to take these stages of the research as given, whereas, as we argue below, they can form the research in very definite ways (See Hutchison, 1981).

So, too, we cannot fully assess a study or series of studies without attempting to assess how the research is *presented*, its *reception*, and how various groups have attempted to *use* the research. Therefore, we need to ask how the research has been used (i) to change, or to buttress, the ideas of teachers, pupils, parents or politicians, as to how education is, or should be, provided and (ii) to argue for recommendations concerning educational policy, or the practice of, say, teachers or parents.

These considerations lead us to try to go beyond what we have called statistically responsible criticism when evaluating research reports. We aim to address a further set of issues to do with the formation and use of the research. We call this approach to evaluating research more broadly 'demystification', as it seems a development of some of the ideas in *Demystifying Social Statistics* (pp 1-4) – but here applied to assessing research reports.

A full assessment of an educational research study must take account of the fact that its ultimate use is generally to persuade people, and to produce support for policies and practices. And this use often depends on the mystique of research, which is often claimed to be 'scientific' and 'objective', because it is based on 'data', especially if these are 'statistical', and if the results are produced by 'computer'.

Thus there is a need to pose, in addition to those questions listed in Section 2.2, the following five questions about any piece of research.

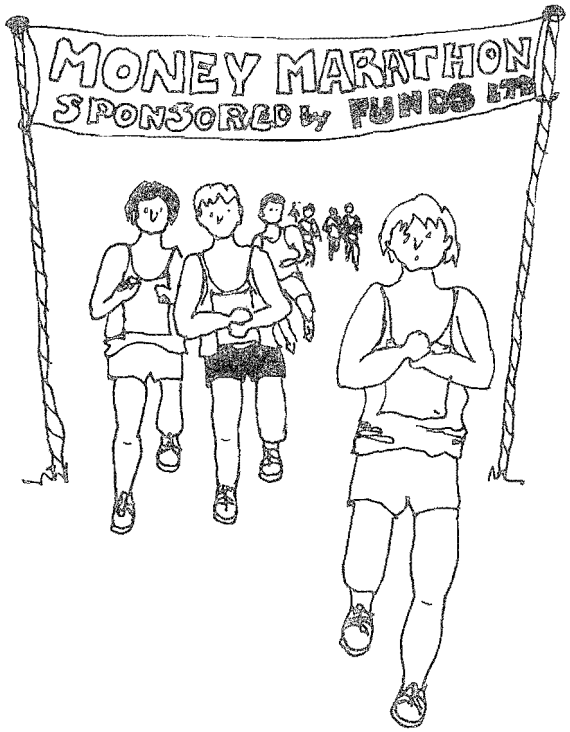
1. Who *commissioned* and *sponsored* the research? Clearly, agencies or persons who commission research, and hence select problems and formulate them in a certain way, as well as sponsors who pay for research and who perhaps grant access, for example, to schools, can play an important role in the formation of research.
2. How was the research problem *conceptualised* and how well did the conceptualisation and the measures used in the study fit together? This is closely related to the issue of construct validity which we mentioned earlier.
3. How was the research *presented* in its printed or other form(s) and how was it *received*?
4. How was the research *used*, and what interests were thus served?
5. Are there any *alternative uses* of the research either within the same or within other conceptual frameworks which could serve other interests?

Each of these questions is considered in turn and linked to the three case studies, particularly *Fifteen Thousand Hours*.

3.2 COMMISSIONING AND SPONSORSHIP

Large-scale educational research, while very cheap compared with the physical and natural sciences, needs larger sums of money in order to carry it out than can be found without approaching an outside sponsor. The major sponsors are SSRC, DES and, until recently, Schools Council, though other bodies such as trusts and ILEA have sponsored educational research. A lot of educational research is funded by central government, but less than 1% of the education budget goes on research. These allocations are dependent on the customer/contractor principle (H.M. Government et al, 1971) which implies that DES are entitled to a major voice in selecting and conceptualising problems to be researched. In practice, this means that research is 'sponsored' within DES for a share of their overall budget on the basis of the current policy interests of specific sections.

Education is not particularly well-favoured within social science research (it gets about 9% of SSRC research funds) and the SSRC only gets about 4% of the total allocation to all research councils. Educational research funded by the SSRC is less subject to the customer/contractor principle although recent changes in the SSRC's structure and the Rothschild enquiry into the SSRC seem designed to bring this principle into play there as well.



— COMPETITION FOR LIMITED FUNDS —

Another way in which the structure of educational research funding influences what is produced as research findings is the competition for limited funds. This, together with the understandable desire on the part of a government department to get value for money, may lead to pressures on a research worker to attempt an over-ambitious study, and to make excessive claims given the resources and the framework of the project. Furthermore, the insecurity in career structures in social research generally means that researchers are tempted to oversell their projects both to sponsors at the beginning and to the media at the end, so that they can get another contract or a 'proper' permanent job.

These general points about funding need to be borne in mind when evaluating educational research. However, it would be naive to suppose that there is a direct link between a particular sponsor and the research findings; educational researchers tend to choose problems and frame approaches to problems in ways which are likely to attract funds but they do not generally try to tailor (or doctor!) their findings to please their sponsors.

Only in one of our three case studies could a direct link be postulated between sponsorship and finished product and that is in the case of Cox and Marks' critique of the National Children's Bureau study. This was carried out under the aegis of the Centre for Policy Studies which was set up by Conservative politicians, and donations to the centre are listed as political donations by companies. Given the enthusiasm of the Conservative party for selective and independent schools, it is perhaps not surprising that Cox and Marks tried to pin labels like 'distorted' and 'biased' on research which showed non-selective schools were generally doing as well as selective ones.

3.3 CONCEPTUALISATION

The world can be conceptualised in a number of ways; put another way, there is more than one answer to the question of what 'things' there are in the world which are possibly relevant to education. What is clear is that without a conceptual framework, a piece of research could never get started: the researchers would not know what they should select as relevant to study, nor how to interpret what they did observe. Preconception – free postures are still found from time to time in research reports even though they have been demolished by philosophers of science: as Popper (1963, Ch.1) says, 'the fact that we can start with pure observations alone, without anything in the nature of a theory is absurd: as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society . . .'

The overall conceptual framework behind any piece must be examined, then, both for which concepts are included and which are left out. In *Fifteen Thousand Hours*, the concepts of school process/ethos are central and 'resources' are not really included in the conceptual framework that is actually used in the study design. In *Progress in Secondary Schools* it was the concept of academic *progress* which was important and not that of academic *standards*. In *'Teaching Styles and Pupil Progress'*, the concepts of formal and informal teaching styles, as adopted by teachers, were dominant and other characteristics of teachers as well as outside influences on teaching style such as the 11+ examination were ignored.

In justifying their exclusion of resources, Rutter et al. cite previous studies and state that 'It

is clear now from many studies in both Britain and the United States that the variations between schools or between local authorities in either financial resources or size of school class show no clear relationships to differences in scholastic attainment' (p.4). In the same way, they justify the use of 'school process' and conclude: 'It is evident that schools differ on a variety of quite different features and there are strong suggestions that these differences may have an important influence on the children's behaviour and scholastic progress. The possibly relevant features include the amount of teaching and degree of academic emphasis, the extent and nature of ability groupings, teacher expectations, styles of teaching and classroom management, the size of the school, patterns of discipline and the characteristics of the overall school climate or atmosphere' (p.18).

Later in the report, the notion of ethos is brought in (pp.55-56) seemingly to substitute for school process. It is characterized as the 'climate of expectations and modes of behaving' of the school. However, the research concentrates on 'particular happenings and behaviours', rather than on 'the more general attitudes which may lie behind them' (p.56), since the researchers were concerned with 'the sorts of actions which teachers and pupils could take to contribute towards the establishment of an ethos which would enable all those in the school to function well' (p.56). This quotation illustrates both the behaviourist conception of ethos, discussed critically, for example, by Young in Tizard et al. (1979, p.31), and by Pring in Wragg et al. (1980), and the sorts of practical recommendations that are likely to be pressed by the researchers.

The word 'ethos' has an everyday meaning: for example, the *Oxford Illustrated Dictionary* (1962) defines it as the 'characteristic spirit of a community, people or system'. Indeed, the claim that the ethos of the school is crucial in influencing pupil attainment, behaviour, etc. has powerful 'common-sense' plausibility, because of this everyday meaning of ethos. However, in reading the research, this appeal to common-sense should be resisted and attention focussed on the meaning given to the concept by the researchers in their definitions and in the way it is measured (see Section 2.5).

Even more striking are the researchers' published claims to do with the alleged lack of influence of 'resources' on the dependent variables. In the conclusions we are told that differences in outcome were not due to 'physical factors' such as size of the school, age of the buildings or the space available. By the next page, it is claimed that the main source of variation between schools in their effects on children does not lie in 'factors like buildings or resources' (p.80). From then on it is assumed that it has been shown that resources do not matter, although no attempt was made to measure them.

And yet, when we consider the indicators of the school process/ethos variable, some at least seem to be dependent on resource allocation to the school (e.g. outings, decorations of classroom, clerical help). Thus, although resources have not been brought into the conceptual framework as such, it could be argued that they have been brought into the study via the way they affect the various facets of school process, and we come back to this when we talk about re-use. It may be too, that there was little variation in resources between the 12 schools in the study and so it would not have been possible to pick up effects of resources in school outcomes. This does not mean that effects would not have been found in samples of schools where resources vary. After all, a recent report by the Inspectorate argues that expenditure cuts will have an impact on educational standards (DES, 1982).

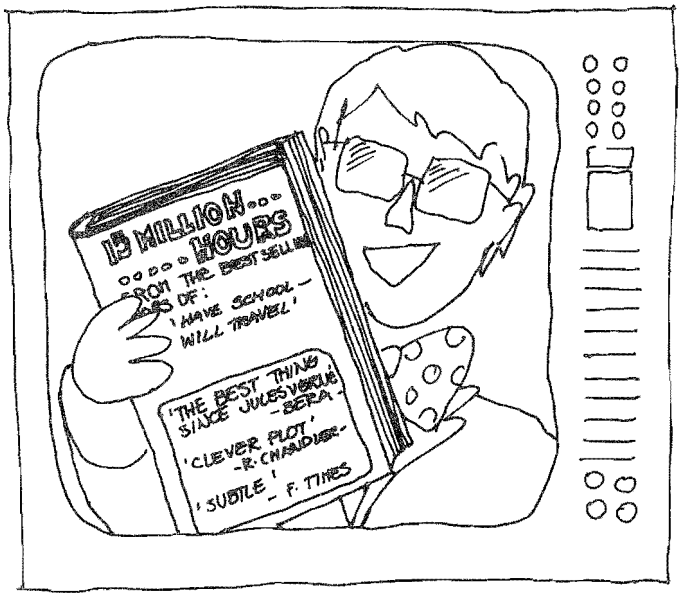
3.4 PRESENTATION AND RECEPTION

Here we look at the form in which research is published, what features of this may serve to mystify its conclusions, and how this mystique is produced. All these are related to the use of the research.

Fifteen Thousand Hours was published amid a fanfare of publicity in March 1979. It was published in the same 'popular' format as *Teaching Styles and Pupil Progress*, by the same publishers. It was claimed to be 'vital reading for all those professionally involved in teaching, and for anyone with a care for the quality of education today' (cover blurb).

There was concern over 'hype' in the reporting of the results and conclusions from both studies. 'Hype' means the promotion of a product beyond its intrinsic merits, in these cases by overselling and by dropping qualifications. This can be evident at a number of stages: for example, in the disjunction between the conclusions reported in the final chapter and the results presented earlier in the book; between cover blurb and the book's contents; between the book's contents and media reporting.

The interaction between the prevailing climates of opinion and hype in presentation can be shown especially effectively in the case of the Bennett research. Bennett brought out his book at a time (1976) when some commentators were mounting a strong attack on 'progressive' methods of teaching. Bennett published his study with an extraordinary amount of publicity, which alleged strong and incontrovertible evidence that progressive teaching was inferior, this despite widely discussed doubts among statisticians about the technique of cluster analysis he was using.



– OVERSELL THEIR PROJECTS –

This discussion may sound excessively negative: what can researchers actually do about presentation? First of all, they can ensure that the conclusions of the research report itself do not show 'hype'. Second, they can attempt to ensure that the publishers, and even the various organs of the mass media do not exaggerate conclusions or present recommendations that are unwarranted. One way of doing this is to try to anticipate later debate, by stating clearly which practical recommendations are, in their view, warranted by the research and which are not. However, we do recognize that it is difficult to prevent hype and that the publicity process gets a momentum of its own which researchers cannot control. And, of course, researchers want their results and conclusions to reach a wide audience and so they understandably court publicity. Ultimately, the only defence against hype is a confident and critical readership.[1]

3.5 USE OF RESEARCH RESULTS AND FINDINGS

We found a number of examples in the press of the use of the presumed relationship between ethos and school effects in *Fifteen Thousand Hours*:

(i) as partial justification for a school closure. Schools with a particularly good ethos should not be amalgamated, even with smaller schools, in the current restructuring, since that would mean the loss of their ethos and style, qualities wanted by parents. (The Education Secretary's justification for not amalgamating Highbury Grove School in North London with Sir Philip Magnus School; reported by John Fairhall, *Education Guardian*, 13 May 1980);

(ii) as a justification for tightening up the hierarchical control structure in a school. In schools where, for instance, exam performance is below average, the head should personally take charge of improving the school ethos by tightening up the authority structure (as might be implied by the term 'tight ship' used in the 'TES', 13 June 1980) – in order to check on punctuality, consistency of discipline and setting homework.

(iii) as something parents were urged to look for when choosing a school. Parents seeking to choose a comprehensive school for their child should ascertain the school's ethos by paying a visit (Maureen O'Connor, 'Education Guardian', 5 August 1980).

These practical conclusions are not justified by the research and this is accepted by at least one member of the team (Peter Mortimore; see Reynolds, 1981).

Consider now the interests served by the conclusions of the research. In the current economic climate the finding that 'resources do not affect school outcomes' will find a ready audience. Education spending increased steadily through the 1960s but has been cut back since the collapse of the economic boom. The current Government is intent on switching resources from education, and other social services, to law and order and to military expenditure. In such a context, the findings can easily be used to justify the switch. Furthermore, the study took place in ILEA, an authority with a high level of spending on education. There have been complaints that perhaps ILEA's performance does not match the high spending level. The findings could be used to justify a cut-back in resources to ILEA.

The implications of the claimed importance of school ethos are more subtle. Fortunately, the researchers' views of these are spelled out in some detail. One quote will hopefully give some indication:

> 'The atmosphere of any particular school will be greatly influenced by the degree to which it functions as a coherent whole, with agreed ways of doing things which are consistent throughout the school and which have the general support of all staff' (p. 192)

This suggests that ethos itself may be interpreted as 'hierarchical', to be laid down by the schools' guardian of ethos (the head?), perhaps in accordance with DES guidelines – or it may be 'democratic', to be developed by the teachers in a school, as they see fit, and in their own way.

Soon after the publication of *Fifteen Thousand Hours*, there appeared to be an official 'line' on the research. For example, a teacher in a London comprehensive tells of the Deputy Head returning from a DES-sponsored conference which discussed the research; the conclusions which the Deputy related to the teachers were that school ethos, and hence performance, were better where:

(a) pupil and teacher punctuality were checked on;
(b) teachers assigned, and checked up on pupils' homework;
(c) senior teachers 'chivvied along' more junior teachers.

This is what might be called a 'hierarchical' interpretation.

Compare that interpretation with the following:

> 'What, after all, do these central chapters say? Better outcomes tend to be associated with schools in which children have their work marked regularly and homework is expected, in which staff know that someone cares enough about this to check that it is done, in which teachers work and plan together, and have the support of interest in their work from the head, through contact with senior staff. Stability of friendship groups, of staffing arrangements, of expectations of how to deal with the day's problems, formalised procedures which mean that teachers know the routines and do not find themselves continually inventing new responses or private codes for passing information – all also tend to be associated with better outcomes . . . Put together under the heading of "the way people deal with people" and "the values which emerge from interpersonal relationships", they constitute, I would suggest, a start in specifying what is meant by "ethos"'. (Graham Bacon, Headmaster in Wragg et al, p.21)

This is obviously a very different type of interpretation of the findings from that of the Deputy Head quoted above. We might call it 'democratic' since it depends for its success on staff working together. All this shows the way in which research such as *Fifteen Thousand Hours* can be quoted in support of very different policies. Either interpretation goes beyond the statistical results. This illustrates that there is still room for argument over policy issues.

We now raise a different question, this time in connection with the National Children's Bureau study, which is linked both to the selection of problems (discussed in Section 3.2) and to the use of research results. Should the research on selective and non-selective schools have been done at all?

The study must be seen in the context of the debate about comprehensives; this debate is perhaps less strident than it was in the 1960s and 1970s and attention has shifted from the state sector to proposals such as the Assisted Places scheme and the future of the public schools. Nevertheless, selection is still an issue and of course some local authorities still operate a selective 11+ system, while others, like Birmingham, are forced to do so because of the existence of a number of voluntary aided foundation schools.

It could be argued that this debate is an ideological one, a debate about means rather than ends. In other words, as many as possible of secondary school children should be in

comprehensive schools for social and political reasons – so that they can mix with children from different backgrounds and so that the majority of children are not stigmatised by being sent to schools which are perceived to be inferior as secondary modern schools tend to be. There are important questions, the argument might continue, to be asked of comprehensive schools about whether and how they achieve these social goals which would have implications for practice in comprehensives, but comparisons of outcomes between different types of secondary school are irrelevant to this debate.

Even if questions about ends are seen to be relevant, it could still be argued that it is (or was, in 1974) too early to evaluate comprehensives by comparing them with selective schools. After all, many comprehensives were, at that time, new organisations trying to work out new methods of working, whereas most grammar schools were well-established with clearly understood functions and good local reputations. Perhaps a fairer comparison could be made now, in 1982, but the research would now be very difficult because only a few local authorities operate a selective system.

Another argument against starting the research is that opponents of a comprehensive system would make political capital out of the results, almost regardless of what the results were. A similar dilemma is described by Cronbach et al. (1980, p. 168); they discuss a case of a proposed project by the United States Civil Rights Commission into school desegregation which was abandoned because it was thought that 'any superiority in achievement of desegregated schools would be tiny at best and that such evidence would encourage opponents of desegregation. Moreover, the average test scores of minority students would presumably fall behind those of the majority, leading once again to disparaging, defeatist headlines'.

Thus, from one point of view, it could be argued that the project had been a mistake. On the other hand, it could be argued that the provision of information in a free society is a necessity, and that it is dangerous to counsel the suppression or even the non-investigation of possibly inconvenient information.

This problem, about the acceptability of certain types of research is not a simple one. We have tried to present both sides of the question to show how the effects of research may extend beyond the immediately obvious, and also to let readers make up their own minds on the subject.

3.6 RE-USE OF RESEARCH RESULTS

The same research teams, or more usually others, may wish to re-analyse a data set, either to test the sensitivity of the conclusions to the particular statistical assumptions and techniques used, or to attempt to interpret the data within a different conceptual framework. An example of the first type is Aitkin's re-analysis of Bennett's data; an example of the second type is given below.

Both types of re-use require the original research team, or whoever 'owns' the data, to grant access to it. The SSRC has a very commendable rule that data from large scale research funded by them should be deposited in the Data Archive at Essex University.[2]

However, the SSRC rule does not always promote effective and prompt access: Bennett did not deposit the data from his study until 1981 although the report was published in 1976. The NCDS data used in *Progress in Secondary Schools* are routinely deposited in the Archive. Unfortunately, the data from *Fifteen Thousand Hours* are not yet available for re-analysis (June 1982), apparently because some of the schools might be recognized and so assurances which the research team gave to the schools about confidentiality would then be breached. We do not find this argument convincing: it may be defensible to withhold some parts of the data to ensure confidentiality but it should be possible to make the bulk of them available to bona fide researchers. Readers might like to consider whether it is right to conduct research, particularly research in politically sensitive areas, in which all the data have to be kept secret.

An example of the second type of re-analysis would involve the development of an alternative conceptual framework to the one used by Rutter et al. We might aim, for example, to study (i) the relationship of resource provision to educational attainment, etc. and (ii) diffferences in the effectiveness of 'hierarchical' and 'democratic' approaches to enhancing school process or ethos.

With (i), the researchers have produced (inadvertently?) what may, on consideration, turn out to be valid indicators for resource provision, at least for the *mix* of resource allocation within the school (e.g. expenditure on clerical help, decoration of classrooms as opposed to outings, or whatever – see their Appendix E), if not for the *total* provision allocated to the school. We could therefore re-analyse the data, using the indicators for resource mix and ethos as separate independent variables. With (ii), we have no indication that the data were produced with the distinction between democratic and hierarchical approaches in mind. Further, when we examine the indicators used for ethos (Appendix E), only one (43; 'decision making') would appear to allow us to begin to distinguish between these two approaches. Thus, (ii) is likely to be impossible to study using these data.

Thus we can see that re-use of the data from a research project depends, first of all, on effective access to it. Re-analysis using different statistical assumptions and techniques is generally possible; whether or not re-analysis using a different conceptual framework is possible will depend on the focuses and constraints of the original conceptual map.

3.7 SUMMARY

This chapter presents a broader basis for the evaluation of educational research bringing in the formative stages which mould the research crucially before data are produced or even before a full research design is developed, and stages of interpretation, presentation and use, which determine the social impact of the study.

This leads us to formulate an additional set of questions which we would argue must be posed in order to 'demystify' the social and political origins of the research and its impact, no matter how 'objective' and uncontroversial its statistical and methodological aspects may seem to be.

End-Notes for Chapter 3

1. It is also worth pointing out that research published in book form tends to escape the scrutiny of professional peers in a way that research reported in academic journals does not. There is something to be said for publishing at least some of the research findings in journals before writing a book although this can lead to frustrating delays.

2. There is also the Scottish Education Data Archive, described by Gray et al. (1979), an important function of which is to be a resource which teachers and others can easily use when wishing to look at aspects of Scottish secondary schools.

CONCLUSION

Quantitative educational research has had an appreciable impact on education and schools. Studies, such as *Teaching Styles and Pupil Progress*, *Progress in Secondary Schools* and *Fifteen Thousand Hours*, have undoubtedly affected the way teachers think about their methods and their pupils, and how effective different types and styles of school organisation are considered to be by administrators and members of the public.

For this reason, we set out in this pamphlet 'to suggest and portray ways in which reports of quantitative educational research can be read more critically', rather than grudgingly accepted – or irrationally rejected – simply because they are based on 'statistics'. To further this aim, we have developed two perspectives, both of which we hope are accessible to those without statistical training, and we have applied them to the three studies mentioned above.

The first perspective, which we have called 'statistical responsibility', is concerned with whether the conclusions are supported by the data and the statistical analysis, and whether they are comprehensibly and fairly presented. This first perspective is the more technical of the two; that is, it focusses on methodological and statistical points. However, we have tried – we hope successfully – to present it in a way that will be understood by lay readers. (Questions about the 'right' statistical technique to use for a given problem are beyond the scope of this pamphlet and could profitably be referred to someone with statistical experience.) In Chapter 2, then, we put forward a checklist of criteria which gives pointers as to where to look in order to decide whether a well-publicised study merits attention, and whether criticism of it is soundly based, or just petty and partisan.

Our checklist of criteria is:

1. construct validity
2. statistical conclusion validity
3. internal validity and 'third variables'
4. external validity
5. levels of analysis
6. explanations of statistical techniques appropriate for the intended audience
7. discrepancies in the care given to interpretation in different parts of the report
8. fabrication of results and deception in presentation

The second perspective, which we have called 'demystification', takes up the theoretical and broadly political questions not addressed by the statistical responsibility checklist. It focusses on the formative stages of the research and the ways it is put to use in society. We argue that educational research needs to be considered as a social product, influenced (consciously or unconsciously) by the preconceptions and interests of those concerned, at all stages from conceptualisation to its impact on educational practice.

In Chapter 3, then, we present an additional set of questions which we consider are necessary to pose about any piece of research. They are:

1. Who commissioned and sponsored the research?
2. How was the research problem conceptualised and how well did the conceptualisation and the measures used in the study fit together?

3. How was the research presented and received?
4. How was the research used?
5. Are there alternative uses of the research?

The two perspectives are certainly different. Though we recognise that some readers will be predisposed to prefer one perspective over the other, we would argue that they are complementary rather than competitive. Our hope is that readers will appreciate the links between the two perspectives (e.g. construct validity, presentation, re-analysis) and will want to use both sets of ideas when evaluating research of this type. Indeed many of the ideas are, we believe, relevant not only to educational research but also to all kinds of quantitative social research.

Some readers may wonder why statisticians are bothering to discuss the educational, social and political issues which are raised in Chapter 3; others may wonder why statisticians claim to speak with special authority about these issues (we don't!). We feel able to speak from both perspectives, perspectives which are essential to the critical consideration of educational research, because we ourselves approach such research from more than one direction. As statisticians, we are concerned that readers appreciate the methodological issues raised by statistical responsibility – and, at the same time, as teachers and researchers at various levels of the educational system, we are 'on the front line' too, and this requires us to be sensitive to the educational, social and political questions.

We are publishing this pamphlet during a period of severe educational cuts. Particular studies can be used by certain interest groups to justify such cuts; we have, for example, argued in Chapter 3 that this was possible in the case of *Fifteen Thousand Hours*. This provides a further practical reason why it seems to us insufficient to concentrate solely on *technical* criticisms of published research while the educational and social implications of the cuts are painfully evident to us as practitioners. Resisting the cuts, we would argue, requires activity on both levels, acting not as isolated statisticians in rarefied professional circles but together with other practitioners in education on common concerns. This is the reason we have emphasised both statistical responsibility and the wider perspectives in writing this pamphlet.

It is socially acceptable to be baffled and bemused by numbers. This is a regrettable state of affairs and means that large sections of the population feel unable to challenge decisions made on the basis of expert technical advice. We hope that people with no more than basic numerical skills – and we would regard the possession of these as a reasonable and important goal for everyone – will be able to apply our ideas when the results from the next educational research project hits the headlines: 'NEW RESEARCH SHOWS . . .' Does it?

GLOSSARY

Here we give brief definitions of the various statistical terms and techniques mentioned in the pamphlet. For further discussion, see the texts listed in the Bibliography or the statistical dictionaries by O'Muircheartaigh and Pitt Francis (1981) and Kendall and Buckland (1975).

* indicates a term which is referred elsewhere in the Glossary

ANALYSIS OF COVARIANCE (ANCOVA) This technique divides up, or partitions, the variance* of a dependent variable* amongst a number of independent variables* and an 'unexplained' category. The independent variables consist of one or more variables* representing groups (sometimes called 'factors') and one or more variables measured on an interval scale* (sometimes called 'covariates'). (The technique is essentially the same as multiple regression*.)

Analyses of covariance were carried out in *Teaching Styles and Pupil Progress* with the variance of post-test scores on a number of attainment tests – the scores obtained at the end of the school year – partitioned into the variance accounted for by teaching style as a factor and the variance accounted for by the 'pre-test' score – the score obtained at the beginning of the year – as a covariate. The aim here was to discover whether teaching style accounted for any variance in post-test score having controlled* for pre-test score. Similar analyses are presented in *Progress in Secondary Schools* and *Fifteen Thousand Hours*.

This type of analysis frequently is accompanied by calculations which show the size of the difference between groups on the dependent variable after controlling for one or more independent variables.

ASSOCIATION See correlation*.

CATEGORICAL VARIABLE See measurement scale*

CLUSTER ANALYSIS The aim of cluster analysis is to put all the elements being studied into two or more groups or clusters on the basis of similarities in their scores on a number of variables*. The clustering method used by Bennett produced 12 clusters of the 468 4th year teachers in the study, using their responses to questions about teaching.

CONTROL The aim of controlling for a variable* is to understand better the actual relationship between two other variables without the relationship being obscured or distorted in some way by the third variable.

We can control for 'third variables'* either at the design stage of the research using random allocation* or matching* or at the analysis stage using techniques such as analysis of covariance* or multiple regression*. All three of our case studies controlled for 'pre-test' score at the analysis stage.

CORRELATION A relationship between two variables* such that a change in one variable

tends to be associated with a change in the other. (For interval scale* variables, correlations generally refer to linear or straight-line relationships.)

A correlation between two variables does not necessarily mean that one causes the other. For example, schools with a *higher* score on a measure of 'ethos' may tend to have *higher* average scores on attainment (a *positive* correlation) yet we could not conclude from this alone that the higher ethos scores caused the higher attainments.

COUNTER-HYPOTHESIS An alternative explanation for the results to that put forward by the researcher. A counter-hypothesis may or may nor be plausible*.

DEPENDENT VARIABLE (OUTCOME VARIABLE) The variable* whose variance* is to be 'explained' or accounted for by one or more independent variables*.

EXPERIMENT (CONTROLLED EXPERIMENT) A design in which the elements (individuals, classrooms etc) are allocated at random* to treatments or conditions.

For example, pupils could be allocated at random to classes within schools in order to produce comparable intakes for each class. This would make it easier to study the effects on attainment of different teaching styles, having controlled* for intake differences at the design stage. However, the use of random allocation in this way can lead to political or ethical difficulties and so controlled experiments are rarely used in education, often being replaced by quasi-experiments*.

Figure G.1 Measurement Scales

	Definition	Examples
Nominal Scale	Values on the scale are categories with no ordering relation between them.	Sex = (M, F) Region of residence in GB.
Ordinal Scale	Scores or values on the scale are rankings or categories which have a defined ordering relative to each other.	Social class when defined by the 6 occupational categories of the Registrar General.
Interval Scale (and **Ratio Scale**)	Equal differences between scores indicate equal differences on the concept being measured. An interval scale allows addition of scores for different individuals and thus the calculation of a mean*.	Attainment test = (0,1,2, ... 100) For these scores to form an interval scale, a difference of 10 marks, between 40 and 50, say, is considered to be equal in attainment to that expressed by the difference between scores of 70 and 80.

INDEPENDENT VARIABLE A variable* which the researcher hopes will account for part of the variance* of a dependent variable*.

INTERVAL SCALE VARIABLE See measurement scale*.

LOG LINEAR MODEL The major aim of this technique is akin to that of analysis of covariance* or multiple regression*. The difference is that the log linear model deals with categorical* dependent variables*; it derives its name from the fact that the mathematical operations take place on the logarithms of the numbers in each cell of a table having two or more 'dimensions'.

In *Fifteen Thousand Hours*, the technique was used to account for variations in attainment, attendance etc. as categorical dependent variables by combinations of intake factors, ethos etc. as categorical independent variables*.

LONGITUDINAL STUDY A design in which measurements are taken on the same individuals (or classrooms or schools) on more than one occasion. All three of our case studies used this design.

MATCHING A method used to make two (or more) groups as alike as possible at the design stage by creating pairs (or triplets etc) of individuals with the same or similar values on relevant independent variables*. This method is not as satisfactory as random allocation* in that the groups may differ not only on exposure to the 'treatment' but also on other, unmeasured, independent variables.

MEAN It gives a measure of central tendency for interval scale* scores. It is the familiar average, the sum of all the observations divided by the number of observations.

MEASUREMENT SCALE The set of values which a variable* can take. Depending on the relations the values are considered to have with each other, we can have three main types in social and educational research; see Figure G.1.

MULTIPLE REGRESSION See analysis of covariance*.

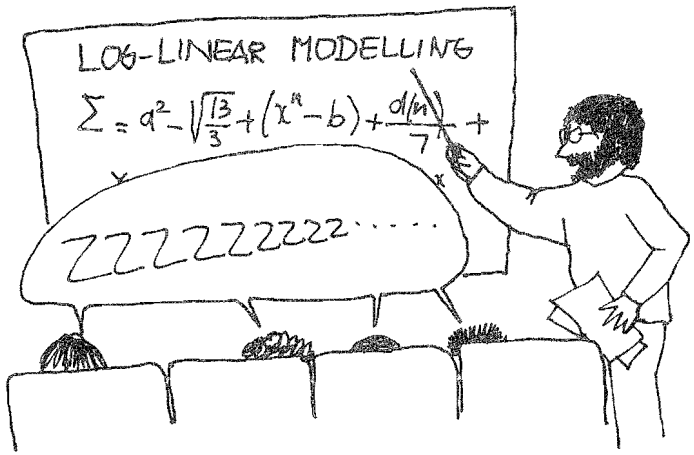
NULL HYPOTHESIS See significance test*.

ORDINAL SCALE VARIABLE See measurement scale*

PLAUSIBLE HYPOTHESIS A hypothesis or explanation for the data which is considered reasonable, given the researcher's (or the critics') theories and the state of knowledge in the field.

POPULATION The set of elements (individuals, schools etc) of interest in a particular study, to be distinguished from the sample of elements which are *actually* studied. For example, all schools in ILEA, all children aged 5-16 in England and Wales.

POWER (of a significance test*) A measure of the ability to reject the null hypothesis* when it is not in fact true.



– EXPLAINING STATISTICAL TECHNIQUES TO A GENERAL AUDIENCE –

QUASI-EXPERIMENT Designs in which random allocation* of elements (individuals, schools etc) to treatments or conditions is *not* used. However, the designs do involve the control* of specified counter-hypotheses* but not of all plausible* counter-hypotheses. There is always the possibility that uncontrolled 'third variables'* may provide the correct explanation for the observed group differences. All three of our case studies can be considered as quasi-experiments.

RANDOM ALLOCATION (RANDOMISATION) A method of allocating elements (individuals, schools etc) to the experiences or treatments of interest using a chance mechanism such as a fair coin or a random number generator. (See experiment*.) It is an ideal way of making treatment groups comparable since we can *expect* that *all* 'third variable'* explanations are controlled* by this device.

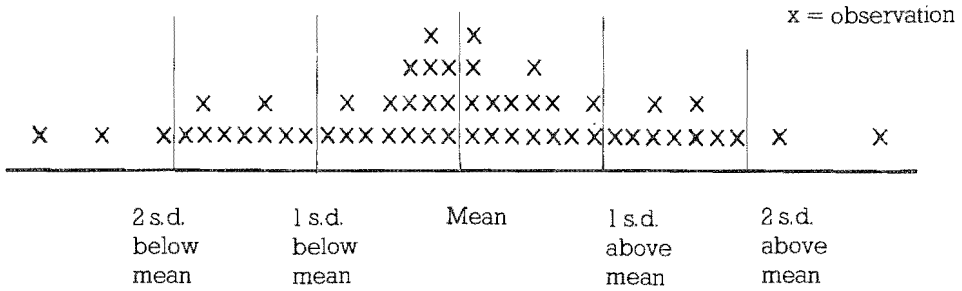
RANDOM SAMPLE A sample drawn from a predefined population* using a random, or chance, mechanism (see random allocation*). A random sample allows 'statistical' inferences to be made from the sample to the population from which it was drawn, but not beyond (except through other types of inference and reasoning). None of our three case studies was based on a random sample of children or schools.

SAMPLING ERROR (SAMPLING VARIATION) The variation in, say, a correlation* calculated from the sample data which would have been observed if all possible samples had been drawn under the chosen sampling scheme.

SIGNIFICANCE TEST (TEST OF STATISTICAL SIGNIFICANCE) A test of a specified statement, or null hypothesis*, about a value or values in the population* e.g. that the population means of two or more groups (say, pupils in selective and non-selective schools) are equal. It assesses how likely it is that the sample actually found would have arisen under that null hypothesis. Such a test is used to reject, or not to reject, the null hypothesis about the population. Significance tests are a way of dealing with the problems of drawing conclusions about the population, given only sample data which is subject to sampling error*.

STANDARD DEVIATION A measure of the 'spread' or 'dispersion' of a set of data, around its mean*. For most variables*, roughly two thirds of the observations or data points will be included in an interval of one standard variation (s.d.) on either side of the mean; see Figure G.2. It is meaningful for interval (or ratio) scale variables* only and can be calculated by taking the square root of the variance*.

Figure G.2 Illustrative Example of Standard Deviation



TENABLE HYPOTHESIS Used to describe a counter-hypothesis* which is both plausible* and supported or corroborated by the researchers' data or other data (see Chapter 2).

THIRD VARIABLE EXPLANATION See control*

VALIDITY Used to denote various aspects of the correctness or applicability of a hypothesis or explanation offered as one of the conclusions of a study (see Chapter 2).

Construct validity refers to the issue of whether the concepts underlying the research have been measured by appropriate variables*.

Internal validity refers to the issue of whether the explanation given is correct for the sample and the context studied, i.e. whether all plausible* counter-hypotheses* have been controlled* for.

External validity refers to the issue of how widely the conclusions of the present study can be generalised to other populations* and contexts.

Statistical conclusion validity refers to the issue of whether sampling error* has been taken account of.

VARIABLE A measure or indicator of a particular concept which can take different values for different members of the sample.

VARIANCE See standard deviation*

BIBLIOGRAPHY

This bibliography is divided into 3 sections. The first consists of the references given in the text. The second is a compilation of critiques of the three studies discussed in this pamphlet. These should not be taken to represent the entire critical response but are intended to enable readers to become further acquainted with the essential arguments of each debate. The third section is a short set of readings on Educational research methods and statistics. Sections 1, 2 and 3 of necessity overlap to some extent.

1. REFERENCES

AITKIN, M.A., BENNETT, S.N. and HESKETH, J. (1981) Teaching styles and pupil progress; a re-analysis. *British Journal of Educational Psychology*, 51, 170-186.

BALDWIN, R.W., (1977) The Dissolution of the Grammar School. In Cox, C.B. and Boyson, R. Eds., *Black Paper, 1977*. London: Temple Smith.

BENNETT, S.N. et al. (1976) *Teaching Styles and pupil progress*. London: Open Books.

BROSS, I. (1960) Statistical Criticism. *Cancer*. 13, 394-400. Reprinted in Tufte, E.R. Ed. (1970) *Quantitative analysis of social problems*. Addison-Wesley. 97-108.

CAMPBELL, D.T. & STANLEY J.C. (1966) *Experimental and Quasi-experimental designs for research*, Chicago: Rand McNally.

COLEMAN, J.S. et al. (1966) *Equality of Educational Opportunity*, Washington, U.S.

COOK, T.D. & CAMPBELL, D.T. (1979) *Quasi-experimentation. Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

COX, C. & MARKS, J. (1980) *Real Concern: an Appraisal of the National Children's Bureau Report on 'Progress in Secondary Schools' by Jane Steedman*. London: Centre for policy studies.

CRONBACH, L.J. et al. (1980) *Towards Reform of Program Evaluation – Aims, Methods and Institutional Arrangements*. San Francisco: Jossey-Bass.

DEPARTMENT OF EDUCATION AND SCIENCE (1982) *Report by Her Majesty's Inspectors on the Effects of Local Authority Expenditure Policies on the Education Service in England - 1981*. London: DES.

EVANS, J. (1982a) After Fifteen Thousand Hours; where do we go from here? *School Organisation*, 2(3), (forthcoming).

EVANS, J. (1982b) Criteria of Validity in Social Research: Exploring the Relationship between Ethnographic and Quantitative Approaches. In Hammersley, M. Ed., *The Ethnography of Schooling*. Driffield, Yorks: Nafferton.

FIELDING, A. (1981) Official Statistics of Education in the United Kingdom: A description of sources and an appraisal. *Review of Public Data Use*, 9, 57-78.

GOLDSTEIN, H. (1977) Sticky Problems of 'creamier' grammars. *The Teacher* 1 April 1977.

GRAY, J. (1981) Towards effective schools: problems and progress in British research. *British Educational Research Journal*, 7(1), 59-69.

GRAY, J., and SATTERLY, D. (1981) Formal or informal? A reassessment of the British evidence. *British Journal of Educational Psychology*, 51, 187-196.

H.M. GOVERNMENT, LORD ROTHSCHILD and COUNCIL FOR SCIENTIFIC POLICY (1971). A Framework for Government Research and Development. HMSO. Cmnd. 4814.

HUTCHISON, D. (1981) The use of statistics in Government decision-making with particular reference the reports of Royal Commissions. *Bias*, 8(2), 179-223.

HUTCHISON, D., PROSSER, H and WEDGE, P. (1979) The Prediction of Educational failure. *Educ. Studies*, 5, 73-82.

IRVINE, J., MILES, I. and EVANS, J. Eds. (1979) *Demystifying Social Statistics* London: Pluto Press.

JENCKS. C., et al. (1972) *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic books.

KENDALL, M.G. and BUCKLAND, W.R. (1975) *A Dictionary of Statistical Terms*. Edinburgh: Oliver and Boyd.

O'MUIRCHEARTAIGH, C. and PITT FRANCIS, D. (1981) *Statistics: a Dictionary of Terms and Ideas*. London. Arrow Books.

PLOWDEN REPORT (1967) *Children and their Primary Schools, Volumes 1 and 2*. Central Advisory Council for Education (England).

POPPER, K. (1963) *Conjectures and Refutations*. London: Routledge and Kegan Paul.

REYNOLDS, D. (1981) School Effectiveness Studies – the emergence of a fledgeling paradigm. *Research Intelligence*. April 1981.

RUTTER, M., MAUGHAN, B., MORTIMORE, P. AND OUSTON, J. with SMITH, A. (1979) *Fifteen Thousand Hours: Secondary Schools and their Effects on Children*. London: Open Books.

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

STEEDMAN, J., FOGELMAN, K. and HUTCHISON, D. (1980) *Real Research: a Rebuttal of Allegations*. London: National Children's Bureau.

TIZARD, B. et al. (1980) *15,000 Hours: A Discussion*. Bedford Way Papers, London University Institute of Education.

WRAGG, E. et al. (1980) *The Rutter Research*. Perspectives, 1. School of Education, University of Exeter.

2. CRITIQUES OF THE THREE CASE STUDIES

(i) of *Teaching Styles and Pupil Progress* by BENNETT et al (1976)

BENNETT, S.N. and JORDAN, J. (1975) A typology of teaching styles in primary schools. *British Journal of Educational Psychology*, 45, 20-28.

GRAY, J. and SATTERLY, D. (1976) A chapter of errors: Teaching Styles and Pupil Progress in retrospect. *Educational Research*, 19, 45-56.

SATTERLY, D. and GRAY, J. (1976) Two statistical problems in classroom research. Unpublished paper; Bristol University School of Education.

BENNETT, S.N. and ENTWISTLE, N.J. (1976) Rite and wrong; a reply to 'a Chapter of Errors'. *Educational Research*, 19(3), 217-222.

GRAY, J. and SATTERLY, D. (1978) Time to learn? *Educational Research*, 20(2), 137-142.

AITKIN, M.A., BENNETT, S.N. and HESKETH, J. (1981) Teaching Styles and Pupil Progress: a re-analysis. *British Journal of Educational Psychology*, 51, 170-186.

GRAY, J. and SATTERLY, D. (1981) Formal or Informal? A reassessment of the British evidence. *British Journal of Educational Psychology*, 51, 187-196.

(ii) of *Fifteen Thousand Hours: Secondary Schools and their Effects on Children* by RUTTER et al (1979)

ACTON, T.A. (1980) Educational criteria of success: Some problems in the work of Rutter, Maughan, Mortimore and Ouston. *Educational Research*, 22, 163-169.
DOE, R. (1980) Second Thoughts on the Rutter ethos. *Times Educational Supplement*, 13th June 1980.
PREECE, P. (1979) Some problems in the analysis of observational data on the effectiveness of schools. *The Rutter Research*, Perspectives, 1, School of Education, Exeter University.
REYNOLDS, D., HARGREAVES, A., and BLACKSTONE, T. (1980) Review Symposium: Fifteen Thousand Hours. *British Journal of the Sociology of Education*, 1(2), 207-219.
TIZARD, B. et al. (1980) *15,000 Hours: A Discussion.* Bedford Way Papers: London University Institute of Education.
HEATH, A. and CLIFFORD, P. (1980) The 70,000 Hours that Rutter left out. *Oxford Review of Education* 6(1), 3-19.
MAUGHAN, B. et al. (1980) Fifteen Thousand Hours: a reply to Heath and Clifford *Oxford Review of Education*, 6(3), 289-303.
HEATH, A. and CLIFFORD, P. (1981) The measurement and explanation of school differences. *Oxford Review of Education*, 7(1), 33-40.
GRAY, J. (1981) Towards effective schools: problems and progress in British research, *British Educational Research Journal*, 7(1), 59-69.
RUTTER, M., MAUGHAN, B., MORTIMORE, P. and OUSTON, J. (1980) Educational criteria of success: a reply to Acton. *Educational Research*, 22, 170-173.

(iii) of *Progress in Secondary Schools* by STEEDMAN, J. (1980).

COX, C. and MARKS, J. (1980) *Real Concern: an Appraisal of the National Children's Bureau report on 'Progress in Secondary Schools' by Jane Steedman* London: Centre for Policy Studies.
STEEDMAN, J., FOGELMAN, K. and HUTCHISON, D. (1980) *Real Research: a rebuttal of Allegations.* London: National Children's Bureau.
LACEY, C. (1981) The comprehensive schools debate. *Times Higher Education Supplement.* 16th January, p.12.
GRAY, J. (1981) Review of 'Progress in Secondary Schools', Jane Steedman, 1980. London National Children's Bureau. *British Educational Research Journal*, 7(1).

3. USEFUL BOOKS AND READINGS IN EDUCATIONAL RESEARCH METHODS AND STATISTICS

BLALOCK, H.M. (1971) *Social Statistics*; 2nd edn. London: McGraw-Hill.
FULLER, M. and LURY, D.A. (1977) *Statistics Workbook.* Oxford: Philip Allan.
GOLDSTEIN, H. (1979) *The Design and Analysis of Longitudinal Studies: their Role in the Measurement of Change.* London: Academic Press.
HARDYCK, C. and PETRINOVITCH, C.F. (1975) *Understanding Research in Social Science: a Practical guide to understanding social and behavioural research.* New York: W.B. Saunders.

IRVINE, J., MILES, I. and EVANS, J. (1979) *Demystifying Social Statistics.* London: Pluto Press.
KALTON, G. (1966) *Introduction to Statistical Ideas for Social Scientists.* London: Chapman and Hall.
KATZER, J.K., COOK, K.H. and CROUCH, W.W. (1978) *Evaluating Information: A guide for Users of social science research* Reading, Mass: Addison-Wesley.
MOORE, D.S. (1976) *Statistics: Concepts and Controversies.* San Francisco: W.K. Freeman.
OPEN UNIVERSITY (1979) *DE304: Research Methods in Education and the Social Sciences*, Blocks 1-8. Milton Keynes: O.U. Press.
OPEN UNIVERSITY (1983) *MDST 242: Statistics in Society*, Blocks A-C. Milton Keynes: O.U. Press.
PLEWIS, I. Ed. (1981) *Publishing School Examination Results: a Discussion.* Bedford Way Papers: London University Institute of Education.
SATTERLY, D. (1981) *Assessment in Schools.* Volume 1 in *Theory and Practice in Education.* Oxford: Basil Blackwell.
STERN, P.C. (1979) *Evaluating Social Science Research.* New York: O.U.P.
TANUR, J. et al. (1972) *Statistics: a Guide to the Unknown.* San Francisco: Holden Day.

Radical Statistics Group was formed in 1975 by statisticians and research workers drawn together by a common concern about the political assumptions and implications of much of their work, and of the actual and potential uses of statistical data and techniques. Membership of the group is open to all those working in or interested in statistics from a politically radical perspective.

Within Radical Statistics are groups with special interests in the political implications of applications of statistics in specific areas such as health, education, race relations and nuclear disarmament. The *Radical Statistics Newsletter* is circulated to all members of the group.

For futher details please contact:

Radical Statistics,
c/o BSSRS,
9 Poland St.,
London W1V 3DG.

Previous Radical Statistics Group publications:

1. Whose priorities? A critique of 'Priorities for health and personal services in England' 1976. 45p plus 20p p&p.
2. Indefence of the NHS. An attack on fee for service payments in medical care. 1972. 50p plus 20p p&p.
3. *RAW(P) deals. A critique of 'Sharing resources for health in England'.* 1978. 25p plus 20p p&p.
4. *Social indicators: for individual or social control? The case of OECD.* 1978. Out of print.
5. *The unofficial guide to official statistics.* 1980; 2nd edition 1981. £1.00 (individual trades union branches, community groups); £2.00 (libraries and public institutions); £5.00 (private institutions and companies); plus 25p p&p.
6. *Britain's Black Population.* Produced by the Runnymede Trust and the Radical Statistics Race Group. Published by Heinemann Books, 1980. £4.95 paperback.
7. *A better start in life? Why perinatal statistics vary in different parts of the country* (audio tape). Produced by Radical Statistics Health Group and Local Radio Workshop. 1980. £1.50 plus 50p p&p. Available on cassette, with notes. £2.00 (individuals, community groups, trades union branches); £5.00 (others).
8. *The Nuclear Numbers Game: understanding the statistics behind the bombs.* 1982. £1.50 plus 35p p&p.
9. *Two statistical methods for detecting health hazards at work.* 1982. 60p plus 25p p&p.

Pamphlets 1, 2, 3, 5, and 9 can be obtained from the Radical Statistics Health Group; booklet 8 can be obatained from the Radical Statistics Nuclear Disarmament Group (both groups at the address given above). Audiotape 7 can be obtained from Local Radio Workshop, 12 Praed Mews, London W2 1QY. (01) 402 7651.

Readers of our pamphlets may also find the following book of interest: *Demystifying Social Statistics.* Edited by John Irvine, Ian Miles, Jeff Evans. Published by Pluto Press. 1979. £4.95 paperback; £9.95 hardback.

Reading between the numbers: a critical guide to educational research

Statistics lend an aura of infallibility to findings from large-scale educational research. In fact, the findings are often used to silence those wishing to speak up for their own legitimate interests.

This pamphlet is written for 'consumers' of educational research, including teachers, students, parents, and administrators. It aims to make them more competent and more confident when evaluating research. Not only does it provide a checklist for assessing the statistical qualities of the research, it also shows how the social and political origins of the research and its impact must be questioned.

A feature of the pamphlet is its focus on three recent studies in education:

* Neville Bennett's *Teaching Styles and Pupil Progress*;
* the National Children's Bureau's *Progress in Secondary Schools*;
* Michael Rutter et al's *Fifteen Thousand Hours*.

The pamphlet points the way to more critical reading of future research and thus aims to contribute to campaigns to improve education.

Radical Statistics Education Group

Radical Statistics
c/o BSSRS,
9 Poland St.,
London W1V 3DG

Publication No. 10
ISBN 0 9502541 9 3

Price £1.25
Postage and Packing 25p