Statistics Without Statisticians - The Case of Education

BY

HARVEY GOLDSTEIN

Institute of Education

University of London

Paper to be read to R.S.S. Social Statistics Section Conference, April 14 1981

Introduction

Out of about 670 academic statisticians listed in the 1980 directory of Academic Statisticians in the United Kingdom, not more than a dozen declare an interest in educational research or the teaching of statistics in schools, and nearly half of these turn up in the Department of Statistics and Computing at the London Institute of Education. This contrasts with about 50 statisticians who state an interest in social science generally and about 120 who refer to an interest in some aspect of medical statistics. A recent visit to my local University bookshop revealed 5 books devoted to introductory educational statistics and none was written by a statistician. Whilst I clearly need to be cautious in making inferences from such a sample, these findings and my own experience after several years in the business of education and educational research, convince me that statisticians have played a very minor role in the development of quantitative educational research and instruction. I believe that this state of affairs does not exist to the same extent in the U.S.A. (and the existence of the new Journal of Educational Statistics supports my view), but I know too little about that or other countries to wish to elaborate. The remarks in the rest of this talk therefore will refer only to the United Kingdom.

It is not my intention to go deeply into the reasons for such a state of affairs, although I think it is a topic which would repay a thorough analysis. I do think however, that the absence of statisticians from such an obvious and legitimate area of interest has quite a lot to do with the early history of educational research, when the predominant discipline which was also numerate was psychology. The psychologists, and more specifically later the psychometricians, appeared to be able to provide the expertise required and I do not think it unfair to say that they subsequently held a monopoly which has only very recently shown signs of disintegrating. To illustrate the kind of thing I mean, let me quote you a reasonably typical recent advertisement for a senior post,

> "Applications are invited for the post of Professor of Educational Measurement in the school of education at ..... Applicants are expected to be well qualified in educational psychology with specialisation in psychometrics"

I do not wish to belittle any of the quantitative developments which have occurred in education, including contributions to factor analysis and much sound educational research. Nevertheless, there is a great deal still to be desired and I intend to illustrate some of the weaknesses which exist. I shall also suggest that statisticians might like to play a larger role in education and I will say where I think there are some interesting problems statisticians could usefully tackle. The examples I use to illustrate these problems are selective and I make no claims for representativeness. They are, however, among the most important and well known pieces of educational research of the last few years, and that seems justification enough.

The Rutter Report

In the 1960's, the majority interpretation of much research into educational attainment, was that the dominating influences were to be found outside the school in a child's home and culture, and that specifically school-based factors had

relatively little effect. By the mid 1970's a reaction had begun to set in against this view (reactions always do set in) and with a conclusion which supported this reaction came the publication of a study (Rutter et al 1979) arguing that schools did differ in their 'outcomes' and that the 'ethos' of a school was an important determinant of educational attainment. The study achieved considerable publicity in the media, initially nearly all of it highly favourable, (see Guardian, 23.3.79, New Statesman, 23.3.79) and the idea that schools should concentrate on obtaining a good 'ethos' seems to have acquired a lasting place in the mythology of education.

The study itself followed up a 'cohort' of approximately 2,000 children from before their entry to secondary school until their first public examinations. For these children, who attended 12 Inner London secondary schools, there were outcome measurements of attendance, delinquency, behaviour and examination results and the basic analysis compared average values for the 12 schools. One set of analyses then studied the extent to which differences between schools could be accounted for by various characteristics of the schools. A second set of analyses, this time using the school rather than the child as the unit of analysis, also looked at the way in which various characteristics of the schools were associated with the outcome measures. As the authors rightly point out, for both types of analyses, causal inferences from such survey data are greatly strengthened if appropriate adjustments are made for pre-existing differences between the intakes to the schools, and this they attempted to do.

A detailed critique of some of the statistical procedures used in this study is given elsewhere (Goldstein, 1980). I want here to make two points and I shall be concerned with the exam results outcome. Firstly, the authors' definition of 'ethos' is very odd. They begin with 46 somewhat mixed variables which measure aspects of classroom organisation etc. and from these choose the 39 which have 'statistically significant' correlations with at least one outcome variable. These are then summed to give a composite score. Hence the definition of 'ethos' is a 'statistical' one and there seems to have been little <u>educational</u> rationale,

even behind the original 46 variables. In fact, for most non-statistical readers, the way in which this analysis is ~~derived~~ described would make it rather difficult to sort out what had really taken place. If it had been clearly stated that the measure of ethos was constructed on the basis of those variables which best predicted examination results etc., I doubt whether it would have had quite so enthusiastic a reception! To my mind, the researchers gave up their responsibility for providing a sound educational basis for defining this concept and fell back on a statistical technique to do if for them. This is not to say, of course, that an educational basis would be easy to formulate, nor that statistical modelling has no part to play, but merely that a measurement which purports to have educational content should have an educational argument to support it. The authors of this report seem not to appreciate this need, and this is an example of the way in which statistical technique can act as an easy substitute for careful and perhaps difficult substantive argument.

The second point concerns the use of different units of analysis. Most of the analyses use the school as a unit, and study relationships using school means. The trouble with this is that the sample size is then only twelve - quite apart from the fact that the sample has certainly not been selected by any random procedure from a recognisable population. The authors carry out numerous significance tests, many of which, unsurprisingly, never reach the 5% level. The authors rely very heavily on P values - a striking example being where they dismiss a large difference in mean examination results between boys schools and girls schools (in favour of the latter) as of 'negligible' importance. The inability to handle properly units ~~of~~ at different levels ~~of~~ of aggregation is also the outstanding feature of my next example.

The Bennett Research

Another reaction in education taking place in the 1970's was that against so called 'informal' teaching methods. In this case the research which seemed to support this was published, again amid a great media fanfare, by Neville Bennett

(1977). From information on a sample of nearly 500 primary school teachers, 36 were categorised as having formal, informal or 'mixed' teaching styles and a study of the attainments of their pupils was carried out. The children were measured at the beginning and end of an academic year and the post-test means compared after adjusting for pre-test differences. The broad conclusions were that in the areas of mathematics, reading and English, children taught by teachers practising 'formal' styles of teaching made significantly more progress than those taught by 'informal' teachers, with children taught by 'mixed' styles teachers in between for English, similar to 'informal' for mathematics and similar to 'formal for reading. One of the peculiarities of the design of this study was that, despite originally having 38 binary variables, the researchers decided to collapse these (via cluster analyses) to a single dimension for the purpose of comparisons. Thus, the contributions of any of the original variables could not be separately assessed since the teachers within each teaching 'style' had similar values on all these. This approach, which is directly opposed to the traditional experimental design where variables are cross classified, is perhaps a reflection of a tendency in educational debates to adopt broad label categories which then form the basis for argument. Interestingly, very little of the critical comment on the Bennett STudy make much of this point. Rather more attention, however, was paid to the way in which the analysis had ignored the sample structure, every child within each style being treated as a member of a simple random sample. In fact the sample was 'clustered' with teachers selected within styles and children within classes. This failure to recognise that the main focus of interest was on teachers or classes as the unit of analysis, led to inappropriate inferences. Together with an amount of arbitrariness involved in assigning teachers to categories on the basis of the cluster analysis, this was the subject of much academic criticism. As is often the case however, shutting the stable door does nothing to catch the horse. In this case the horse carried the message of a more 'traditional' approach to teaching and can still be seen contentedly devouring large areas of informal teaching in corners of many local educational authorities.

This particular story, however, looks as if it may have a happy ending, or at least one which undoes some of the damage. In 1979, the SSRC ~~which had funded the original research~~ agreed, through its Statistics Committee, to provide modest funds for a reanalysis, to be carried out jointly by Neville Bennett and Murray Aitkin who had recently started the Centre for Applied Statistics at Lancaster University. The results of that reanalysis are due out in various papers this year (Aitkin & Bennett, 1981), and to cut a long story short, show that many of Bennett's original conclusions do not remain. In particular, once the between-teacher variation is incorporated into the analysis model, most of the comparisons are non-significant and an alternative (latent class) clustering method actually reverses some of the trends. Thus, for example, for reading the 'informal' style teachers have the highest average progress score.

Thus, by taking explicit account of the sample structure Aitkin demonstrated how important was the large variation between teachers within styles, and was also able to point to some of the weaknesses of the original cluster analysis assignment and to show how modified assingments affected the final results. Needless to say, had Murray Aitkin not been at Lancaster, and indeed had not the SSRC Statistics Committee had the foresight to create the original Professorial Fellowship which attracted him there, we would almost surely not have had such an analysis. The lesson, I hope, is obvious.

Procrustean Assessment

My final example is in some ways quite opposite to the first two. Whereas both of those, especially the latter, could benefit from the application of more powerful statistical tools, this example suffers from a surfeit of them.

Recent developments (see e.g. Bartholomew, 1980) have provided models for the factor analysis of binary variables. Thus, an 'individuals' responses to a set of yes/no questions can be related, via a linking function, such as the logit, to a linear function of one or more 'factors'. A simple version of such a model, known as the Rasch Model (see e.g. Wright, 1977) has been used for some time now to describe

the responses of individuals to the items of mental tests, e.g. a test of attainment in mathematics. The model itself can be written simply as

$$\text{logit } P_{ij} = \alpha_i + \beta_j$$

where P_{ij} is the probability that individual i responds correctly to test item j. α_i is interpreted at the 'ability' of an individual and β_j the facility of the item. This model implies, among other things, that the relative 'facilities' of the items are the same for all individuals whatever their ability - a highly questionable assumption in an educational system in which diversity of curriculum is the order of the day. Nevertheless, this model, by about 1977, had been adopted by and large as the mainstay of the newly set up Government Assessment of Performance Unit (APU) which is now engaged in extensive national monitoring of school attainments, with two reports already published (APU, 1980). Some account of this is given in Goldstein (1979), and it is now fair to say that the APU itself is not as enamoured as it was originally of this model, nor do its original proponents, the National Foundation for Educational Research, now appear to have much faith in it. (See Times Educational Supplement 19-4-1).

I really want to make two points. Firstly, if true, the model holds out considerable hope of providing a simple means of making valid comparisons across curriculum and indeed also of making valid absolute comparisons across time - despite the changing relevance of test items and use of different tests. This promise must have seemed very attractive to a unit, one of whose remits was to say something about change over time. Secondly, the statistical basis of the model, while familiar to statisticians, was almost entirely opaque to the educationalists associated with the APU. Thus, a degree of mystification was involved, and the non-statisticians were not really able to query the model's assumptions, nor does there every seem to have been much attempt on the part of its proponents to explain what they really were.

To my mind this is a beautiful example of a procrustean bed, the ancient Greek robber who chopped down the limbs of those travellers too large for his bed and stretched those too small. That story is of course well known, but lesser known is the sequel which is that when Procrustes had collected sufficient cases he wrote a

paper in which he proved that every traveller was exactly the same size!

One final comment on this I would like to make and in doing so to take issue with the view expressed by David Bartholomew (1980). In the binary latent-trait models such as Rasch, or those described by Bartholomew, the use of different linking functions will in general lead to different parameter estimates (Goldstein 1980). Indeed, the linking function plays a crucial role in defining the scale properties of the individual ability! Bartholomew seems to defend the choice of symmetrical functions such as the logit on the grounds that it 'reduces the options to one' and otherwise we would need 'to find some extraneous grounds for preferring a particular form.'

This seems to me to be both dangerous and a counsel of despair and really to be used only when all else has failed. Indeed, if we cannot appeal to educational (or other substantive) criteria to determine scale properties (and hence among other things the rank order of individual's parameter estimates) then a justification on the grounds of mathematical simplicity seems very weak indeed, as well as arbitrary. If this is really the best we can do, then I wonder whether we should really be using such models at all and I begin to worry that if we do, some of our educational colleagues may smell a rat and declare a plague on all our works - which would be a pity.

A Role for Statisticians?

It should be fairly obvious that I think there could be a more fruitful involvement of statisticians in education. That is not solely, however, a question of more statisticians or statistical advice being made available, but of statisticians developing long-term commitments to education. The medical model may be useful, where one finds a well defined and intellectually active field of medical statistics which encompasses both technically competent statisticians with a commitment to medicine and qualified medical practitioners with an understanding of statistics. To an extent, some statisticians are already involved in education, for example, central government, especially the D.E.S. employs statisticians to help design policy, collect and analyse data. Despite doing a competent job however, they

seem to have too little contact in their professional life with academic and research statisticians in order to have a significant effect outside their immediate departmental requirements. In local education authorities, there is a sprinkling of statisticians, but these are typically isolated. There are notable and enlightened exceptions in the Inner London Education Authority Statistics and Research Unit which has a respectable tradition of useful and often innovative research. There are, also, statisticians associated with individual research units and projects and some good educational statistical work is also done in some of the research units attached to public examination boards such as the A.E.B. Of course, there is also the extremely important work of the Scottish Education Department in co-operation with academics, which Dr Wishart will mention. There is, however, still a basic lack in the academic area. Without a sizeable number of academic statisticians actively pursuing research in educational statistics, not merely as a passing interest to obtain data for students or to try out a new method of estimation, but with a commitment to the pursuit of educational knowledge, then the necessary cohesion and development will be lacking.

There is certainly no lack of interesting problems to be explored; the Bennett study is just one example of a problem which succeeded in attracting a statistician and produced worthwhile research. The development of models for a hierarchically structured system tackled by this research is of general interest, and in education is of central importance for much research. For example, a part of the present debate about the publication of secondary schools' examination results is concerned with adjusting results for school intake differences. This involves quite subtle differences of interpretation depending on whether the school or the individual child is considered as the 'unit of analysis' and there are several interesting statistical issues here which will need attention. Central government seems to have become increasingly involved in research, not only in the APU but also, for example, in the survey activity of the Inspectorate. Provided that the present government does not pursue its obsession with short sighted financial savings to the point where there is little government research activity left at

all, there is an important area here where the advice and involvement of statisticians would be useful and indeed I believe welcomed. By the same token, local education authorities are continuously collecting statistical information which is all too often insufficiently analysed, and where there may be no professional advice available to ensure efficient procedures. To take one example, there is a great deal of educational testing taking place among LEA's now, especially in the primary school and particularly in reading and mathematics. In most cases, the data generated are simply filed away, or a few simple descriptive statistics produced. Little attempt is made to use the results, for example, to provide local norms, or to carry out analyses designed to measure the suitability of the tests for the local population and perhaps develop new instruments. At the Institute of Education we are currently working with one outer London education authority along such lines, but I know of no such work elsewhere in England and Wales. Incidentally our experience with this project has shown us now important it is that such co-operation should involve an educational commitment by statisticians if a good working relationship is to be achieved and really fruitful results produced. The Scottish model of co-operation is another example of a useful one which might be followed elsewhere.

Finally, let me say that while I am reasonably sure about what I would like to see, I am not sure I know just how to achieve it. At a time of university expansion it would be appropriate to see the establishment of statistical departments within university schools of education. As it is I think that the academic community will move slowly if at all, but even a slow increase of interest would give me great pleasure.

REFERENCES

1.	RUTTER, M Et al	(1979)	Fifteen Thousand Hours London, Open Books
2.	BENNETT, N	(1976)	Teaching Styles and Pupil Progress London, Open Books
3.	WRIGHT, B D	(1977)	Misunderstanding the Rasch Model J. of Educ. Meas. 14, 219-225
4.	GOLDSTEIN, H	(1979)	Consequences of Using the Rasch Model for Educational Research. Br. Ed. Res. J. 5, 211-220
5.	AITKIN, M A, BENNETT, S N & MESKETH, J	(1981)	Teaching Styles and Pupil Progress: A Reanalysis Br. J. Ed. Psych (To Appear)