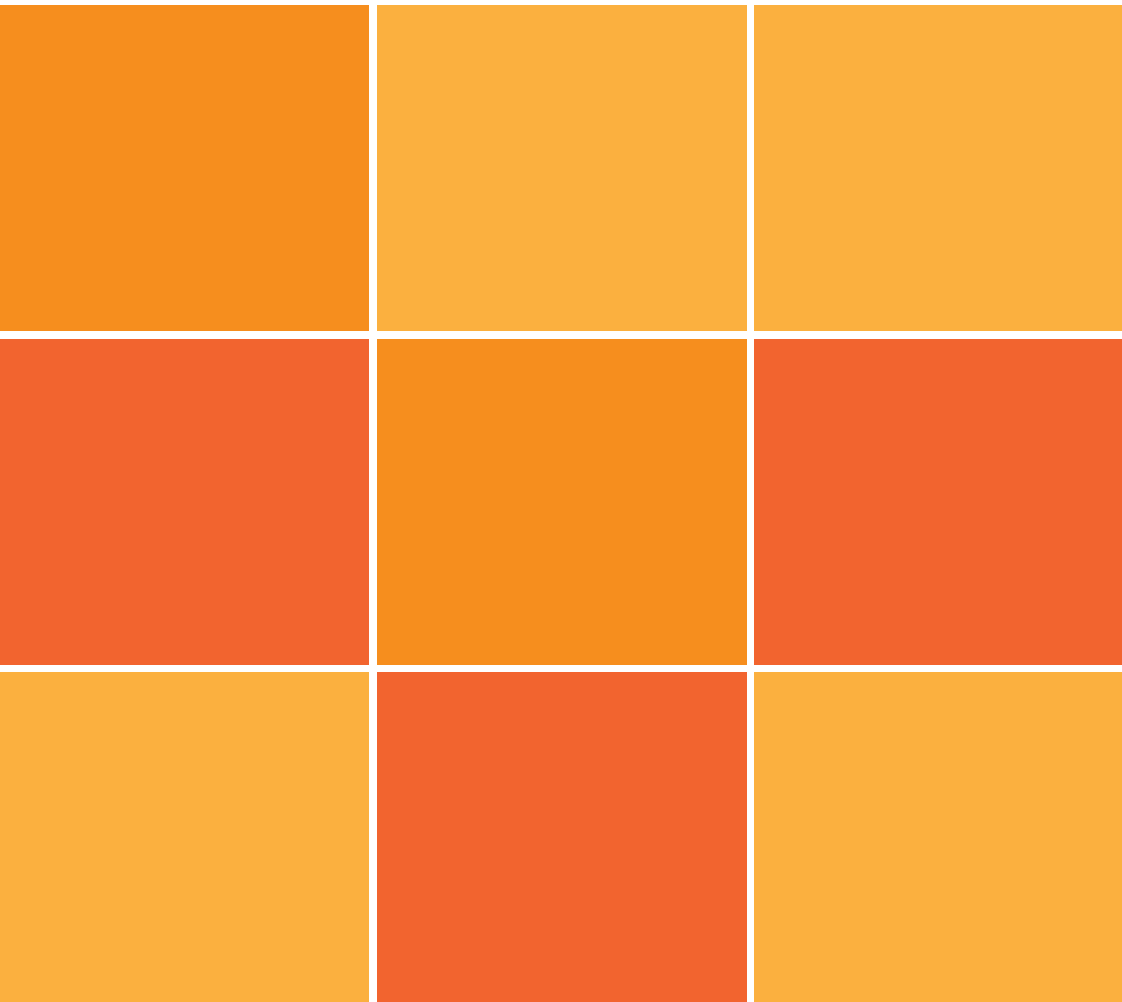
Leading education and social research
Institute of Education
University of London

Predictions, explanations and causal effects from longitudinal data

A professorial lecture by Ian Plewis



Predictions, explanations and causal effects from longitudinal data

Ian Plewis

First published in 2007 by the Institute of Education, University of London,
20 Bedford Way, London WC1H 0AL
www.ioe.ac.uk/publications

British Library Cataloguing in Publication Data:
A catalogue record for this publication is available from the British Library

ISBN 978 0 85473 775 8

Typeset by Keystroke, 28 High Street, Tettenhall, Wolverhampton

Printed by TO COME

Institute of Education • University of London

Predictions, explanations and causal effects from longitudinal data

Ian Plewis

Professor of Longitudinal Research Methods in Education

Based on a Professorial Lecture delivered at the Institute of Education,
University of London on 31 January 2007

Professor Ian Plewis

Predictions, explanations and causal effects from longitudinal data

In 1861, William Farr – one of the founding fathers of modern epidemiology – wrote:

> Again I must repeat my objections to intermingling Causation with Statistics. . . . The statistician has nothing to do with causation; he is almost certain in the present state of knowledge to err. . . . You complain that your report would be dry. The dryer the better. Statistics should be the dryest of all reading.
>
> (Quoted by Diamond and Stone, 1981: 70)

Some of you might indeed think that statistics, and perhaps also statisticians, are dry, although I'm sure that others might see this as a compliment rather than as a condemnation. What might surprise many of you is that the recipient of the letter from which the quotation is taken – and the person wishing to draw causal conclusions from statistical material – was Florence Nightingale. For many people Florence Nightingale was, of course, the 'Lady with the Lamp' but some of us think of her as much by another sobriquet: 'Passionate Statistician'. I shall return to Farr, to his contemporaries and his successors, and to the contested position of causality within the social sciences in due course.

Florence Nightingale

This lecture is about the value of data generated from what are known collectively as longitudinal studies and the ways in which these data can be used to describe and to explain the world that interests researchers in the social sciences. I reflect on some of the studies that I have been involved with in my research career, the great bulk of which has been spent here at the Institute. I do so from the perspective of someone interested in the combination of substantive research questions, research design and statistical analysis to produce better descriptions and understandings of the world.

If we measure the same cases more than once, ideally many times, then we give ourselves the opportunity to find out more about the world than we can from measuring cases on just one occasion. The value of this longitudinal approach, as it is called, is now widely recognised by empirically-minded social scientists who have at their disposal a burgeoning collection of longitudinal resources. The position today is very different from the one that I discovered 30 years ago as I started work on a Social Science Research Council (SSRC) Fellowship on the statistical problems of longitudinal studies. There was no

British Household Panel Survey (BHPS) then; the census-based Longitudinal Study (the LS) had just started and was still only cross-sectional; the 1970 birth cohort study had barely begun. Instead, there were a few small-scale studies, mostly in psychology, one of which was the prescient Children's Centre study initiated by Jack Tizard at the Thomas Coram Research Unit (TCRU), which I had been working on for the previous few years. And, very importantly, there were two cohort studies – the 1946 National Survey of Health and Development, and the 1958 National Child Development Study or NCDS.

Not only were there rather few longitudinal resources for social scientists at that time – and it is worth remembering that although NCDS is a major resource for social scientists now, it started out life as a cross-sectional investigation into the causes of perinatal mortality – there was also little knowledge within the social science community about how to analyse longitudinal data. It was, I think, that lack of knowledge and experience that led the SSRC to invest in my Fellowship. And so, Harvey Goldstein – who had recently arrived at the Institute from the National Children's Bureau where he had been closely involved with the NCDS – and I put together a case that persuaded SSRC to fund me for what turned out to be three days a week for four years.

I would like to think that the SSRC invested wisely but if they did then it turned out to be a long-term investment, because the 1980s were dark years for the social sciences. To quote from a recent publication:

> However, in 1979 the climate turned colder still with the election of the Government led by Margaret Thatcher. Despite her personal interest in scientific research, previous Conservative Governments had made clear their opposition to establishing any national funding body for the social sciences. This scepticism about the value of social science research manifested itself in 1981 when the then Secretary of State for Education and Science (but more importantly Mrs Thatcher's intellectual guru) Sir Keith Joseph, announced that he had asked Lord Rothschild to conduct an independent review into the scale and character of the work of the SSRC.
>
> (Economic and Social Research Council 2006: 18)

The SSRC, perhaps against the odds, did survive, losing its name but not its identity. However, the travails of the research council were accompanied by a move away from quantitative social science and especially from longitudinal research. Thus, apart from the LS (what is now the Office for National Statistics Longitudinal Study or ONS/LS), no major UK longitudinal study started between 1970 (BCS70) and 1991 (BHPS).

How different the situation is today: the 1946 cohort is reaching retirement age and continues to provide important data for epidemiologists and social scientists; we have 14 waves of data from the BHPS; the ONS/LS now has data from four linked censuses; and, soon, the very large UK Longitudinal Household Survey will go into the field. Moreover, even the long-running General Household Survey from ONS has changed from a repeated cross-sectional design to one with a rotating panel design. And within the Centre for Longitudinal Studies (CLS) here at the Institute, the NCDS continues to thrive, BCS70 was rescued from an uncertain future by John Fox and John Bynner at City University and is now on a sound footing, and the Millennium Cohort Study, carefully nurtured by Heather Joshi from infancy and the first British cohort study planned from the outset as a resource primarily for social scientists, is embarking on the fourth wave of data collection. Not only are there many more resources than there were in 1977, there is also more expertise in the organisation of longitudinal databases and the associated statistical analyses. Researchers need no longer be frightened by the structure and analysis of longitudinal datasets.

What can social researchers do with longitudinal data that we are not able to do when we have only cross-sectional data? In other words, what are the advantages of repeatedly measuring the same cases over time – sometimes measuring the same variable more than once, sometimes measuring different variables – over measurements taken at just one occasion? These advantages can, I think, be placed under four broad headings: prediction, description, causation and explanation. As we shall see, there is some overlap between them, but the aims and activities under each heading are sufficiently different to warrant their separation. These four aspects of the longitudinal approach form the basis of this lecture. My intention is to set out some of the many different ways in which

longitudinal data can be used, and to give you a flavour of the methodological challenges that the studies present and how statisticians attempt to overcome these challenges.

Prediction

Statisticians are not fortune tellers but statistical techniques are used to make predictions. These predictions are probability statements about what can be expected to happen in the future given what we know about the present and the past. Some predictions, for example about climate change, are based on analyses of time series data and this is a rather specialised branch of statistics that I do not consider tonight. Other questions require longitudinal data that are collected over a long time span, for example: what are the chances that someone exposed to a particular set of circumstances and experiences early in their lives will find themselves with no educational qualifications in early adulthood? It is for these kinds of questions that cohort studies come into their own but the answers that they provide can bring with them a further set of questions of a more methodological nature.

The opportunities provided by studies like NCDS and BCS70 to relate adult outcomes to events and circumstances during childhood has generated many insightful analyses by colleagues within CLS, both past and present, as well those within the Centre for the Economics of Education and other parts of the Bedford Group, the ESRC Human Capability and Resilience Network, and elsewhere. These researchers have generally found continuities between childhood and adulthood: to give just one illustration from many, parental interest in their child's education is a predictor of economic success in later life (Kuh *et al.*, 1997). The finding is interesting but it begs many questions, some of which we return to later. The question that I want to address here, necessarily rather briefly, is one that perhaps receives insufficient attention: just how well can we predict adult outcomes from childhood circumstances? Or to put this slightly differently: are the predictions sufficiently accurate that policy makers might sensibly base preventative actions and interventions on them? This question

arose from work on a large-scale evaluation project, the National Evaluation of the Children's Fund, completed last year (Edwards *et al.*, 2006). Although the project suffered, as many evaluations of government programmes do, from changes in the national and departmental political climate, it did enable the team here at the Institute to think about the value of prediction in terms of the ways in which services might be targeted at specific groups. The following conundrum is not, of course, a new one: is it better to target services at those most 'in need' of them, or most 'at risk' of being disadvantaged without them, or to provide them universally, to achieve the policy goal of reducing childhood and later disadvantage?

For the cohort of children born in 1958, I show (Plewis, in preparation) that a set of circumstances and events from their early lives – whether or not they were brought up in social housing, their birth weight, whether or not they had a spell brought up in care, their mother's age at birth – are all associated with whether or not they will have any educational qualifications at age 23. This suggests that it might be worth targeting services at a group of children with some or all of these characteristics – living in social housing, low birth weight, a spell in care, a young mother at birth. On the other hand, the predictions, based on fitting a logistic regression model (the estimates from which are given in Table 1) and the associated receiver operating characteristic or ROC curve (Swets *et al.*, 2000), although better than chance, are not very good. So, if we were presented with a pair of children one of whom we know will and one of whom we know will not end up with qualifications, then decisions about targeting that are based on this statistical prediction rule would be accurate for two such pairs out of three – certainly an improvement over the chance success rate of 50 per cent but not a great improvement.

We could do better if we were to include variables measured at later ages among our predictors. It is widely believed, however, that the key to reducing disadvantage is to intervene early and so the most useful statistical prediction rule is likely to be one that is based on measures taken early in life. We need to consider whether the benefits of reaching some but not all of the group 'at risk' outweigh the costs of providing services to people who, on the face of it, are not in need of them. There are many other social policy issues around targeting that

Table 1 Predictive logistic regression model estimates

Explanatory variable	Estimate (S.E.)
Birthweight (reciprocal)	125 (16)
Mother's age	–0.013 (0.005)
Mother's age squared	0.002 (0.001)
Social housing	1.2 (0.15)
In care	1.0 (0.17)
Grandfather's social class	0.30 (0.030)
Social housing * grandfather's social class	–0.14 (0.044)
Model fit: $\chi^2 = 638$ (7df); n = 8518	

I cannot go into here, but the important message from this example is that statistical modelling of longitudinal data from cohort studies enables us to think about prediction in terms of policy, in particular for resource allocation. What is also important here is that we are concerned more with our ability to predict accurately and less to explain what lies behind the associations – in other words, the processes that lead to one person gaining qualifications and the other not are not so relevant here. If we can show that a statistical prediction rule is reproducible across cohorts, despite possible changes in the marginal probabilities for the outcome and its correlates, then we have a tool that is potentially very applicable even if it does not provide us with all the insights we might, in other circumstances, be searching for.

We can use statistical prediction rules to draw out implications for public policy, but we can also use them in a much more methodological context. One of the problems faced by all longitudinal studies, but especially by long-running cohort studies, is that it is not usually possible to measure each member of the selected sample at every chosen occasion. Sample members cannot always be located, they cannot necessarily be contacted even when they are located and they sometimes choose not to participate even when they are contacted. This non-response has potentially serious implications for substantive conclusions drawn from analyses based just on the measured sample because all the research evidence leads us to conclude that non-responding sample members will be

systematically different from those who respond. If we were able to use the information we have about non-respondents to predict their subsequent response behaviour then we might be able to develop interventions that could reduce future levels of non-response. One of the advantages of longitudinal studies is that we often know something about non-respondents at a particular wave from their responses to earlier waves. Hawkes and Plewis (2006) took advantage of this for a series of analyses of the predictors of non-response for successive waves in NCDS and found that non-respondents do indeed differ from respondents, that wave (or temporary) non-respondents are different from those lost from the sample permanently, and that non-contacts are different from refusals. These different respondent groups are not, however, widely separated by the set of predictor variables that are most strongly related to non-response and so the potential for a targeted intervention is correspondingly reduced. Further investigations are needed here.

These two examples of using longitudinal data for prediction bring out two rather contrasting points. The first is that future circumstances and behaviours can be predicted from current and past measures at a level that is much better than just tossing a coin. The second is that it is not easy to make accurate predictions of the future, at least not given our current level of knowledge, and our ability accurately to measure phenomena of interest. This difficulty is not, of course, a weakness of the longitudinal approach nor does it necessarily lead to uninteresting findings. It shows, for instance, that some people are sufficiently resilient to overcome early disadvantage whereas others slip back despite the most propitious of starts.

Description

Let us return to William Farr's comment to Florence Nightingale that introduced this lecture. It places Farr firmly within the camp of those who believe that statistical analysis ends with description. Although I do not share this view, I do believe that it is a mistake to regard description (together with the statistical inferences that are usually needed to go from a sample to the population)

William Farr

just as a necessary prelude to the more exciting tasks of establishing causes and generating explanations. And longitudinal data provide us with the opportunities for more sophisticated descriptions than cross-sectional data do. One illustration of this comes from the distinction in educational research between the concepts of attainment at a particular point in time and progress over time or with age.

The study that taught me most of what I know about educational research was one directed by Barbara Tizard at the Thomas Coram Research Unit in the 1980s and known as the Infant School Project (Tizard *et al.*, 1988). The research was motivated by a concern that is still with us today: that children of African-Caribbean origin were not doing well at school. We followed a group of young children from their entry from nursery class into 33 inner London infant schools – each school having intakes of both African-Caribbean and white majority

children – up to the end of infant school (now Year 2) and then up to their final year in primary school (now Year 6). It was an unusual study at that time because quantitative data were collected from and about the children from their mothers and from their teachers, and about their schools including systematic observations of the children's behaviour at school. One of the more important findings from the study was that the African-Caribbean girls had higher attainments than the other three groups in reading and writing (although not in maths) at the end of infant school, and they made considerably more academic progress than African-Caribbean boys throughout primary school. One lesson from this finding is that a full understanding of ethnic differences in educational progress should take gender differences into account as well. In statistical terms, there are important interactions between ethnic group and gender.

Innovative as I think the Infant School Project was, its sample of 33 schools was really too small for the detailed analyses that would have qualified it as a full description of the relative progress of different ethnic and gender groups. We are, however, in a much stronger position now to monitor these kinds of differences in attainment and progress. The development, by the Department for Education and Skills (DfES), of the National Pupil Database (NPD) and its links to PLASC, the pupil level annual schools census, provides exciting opportunities for educational researchers with good quantitative skills to look in much more detail at educational inequalities and how they are changing with time. Twenty years on from the TCRU study, have the relative positions of the four groups of interest changed? At the end of Key Stage 2 for 2004/5, we find that the attainment gap between white and African-Caribbean pupils varies between just under one-fifth of a standard deviation unit for English to over one quarter of a standard deviation for maths and science. But, for boys, the differences between the two ethnic groups are consistently larger than they are for girls, although they are rather less marked than they were in the 1980s. Between Key Stages 1 and 2 we find that white pupils make more progress than African-Caribbean pupils and again the ethnic differences in progress are more marked for boys. Figure 1 shows differences in progress in English for all the important ethnic-gender groups relative to white boys. It shows that all groups are making more progress than white boys apart from the black Caribbean boys and this

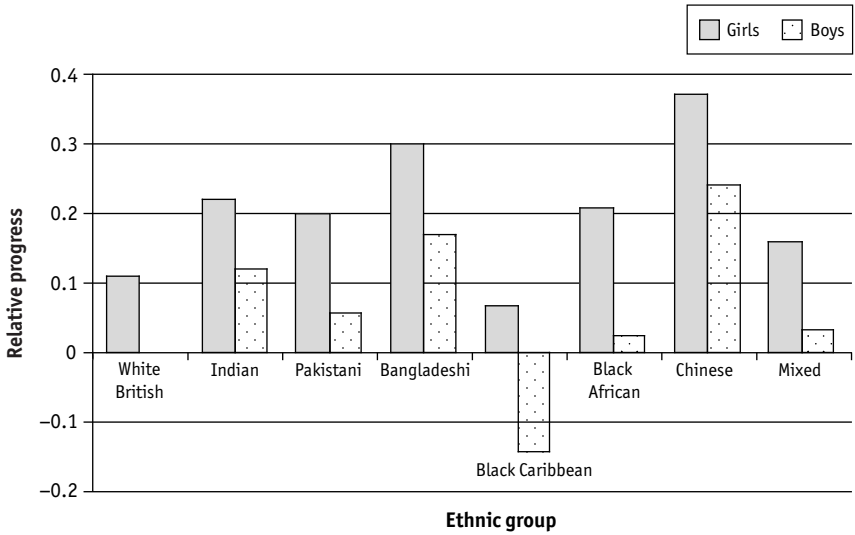


Figure 1 Progress in English (SD units), KS1 to KS2 by ethnic group relative to white boys
Source: DfES National Pupil Database, 2004/5

provides a further illustration of the national story that is worryingly similar to the London story of the 1980s, one that is told in more detail in Plewis (1991). But it also illustrates just how important it is not to create undifferentiated categories of pupils from different ethnic groups and this is where the NPD is so valuable. For example, we see from Figure 1 that pupils with a Bangladeshi background make more progress than white pupils between years 2 and 6 in English, and the same is true for pupils with a black African background.

The size of the National Pupil Database makes it possible to explore some interesting interactions between different correlates of attainment and progress that would remain hidden in smaller samples. Another of its advantages is that it gives the school attended and postcode of residence for each pupil as their school careers unfold. Hence, we can answer interesting questions about school differences in progress, about how differences between local areas combine with differences between schools in their associations with attainments and progress, the association between pupils changing schools and their subsequent (and previous) progress, whether differences between boys and girls in progress

vary across schools, and so on. We can answer these kinds of questions, and hence obtain a much more profound description of difference and inequality, because of one of the most important developments in applied statistics in the last 25 years: the statistical and computing advances that come under the heading of multilevel modelling. These developments owe a great deal to the work of Harvey Goldstein, Jon Rasbash and their colleagues at the Centre for Multilevel Modelling, for so long here at the Institute and now at the University of Bristol. The methods have transformed the analysis of large and complex datasets and offer a lot to analysts of longitudinal data.

Let me illustrate this with just one example from the analysis of progress in English between Key Stages 1 and 2 using the NPD. In addition to the ethnic group and gender differences already discussed, we find that the relation between attainments at these two points in time is complex, with a number of non-linear and interaction terms needed properly to represent it. We also find that the relation between KS1 and KS2 assessments varies from school to school (which, incidentally, is one of the reasons why any kind of school league table is likely to be misleading). In addition, the average difference between boys and girls in progress varies considerably from school to school: favouring boys by 0.05 SD units at one extreme to favouring girls by 0.27 SD units at the other.

There are in fact two ways of analysing differential progress of this kind. The first is by estimating group differences in test scores at the end of the period conditional on pupils' scores on tests taken on entry to reception: this is often known as the regression approach and is the method used for the examples given here. The second approach fits growth curves to each pupil's test scores and examines differences between groups in a multilevel framework of measurement occasions nested within cases (Plewis, 1996). There has been a long debate in the methodological literature about the relative merits of the two approaches, but it is now widely accepted that each is useful in that they answer rather different questions.

Prediction and description are relatively uncontroversial activities for statisticians. The trouble we take properly to describe the world establishes social and educational facts. But these facts do lead us to unanswered questions that we

would hope, despite Farr's words, to pursue further. Why are African-Caribbean boys falling behind in academic subjects? Why were children born to young mothers at greater risk of not obtaining educational qualifications? Can the fact that sex differences in attainment in English widen in some primary schools and narrow in others be related to school policies and classroom teachers' teaching methods? To get anywhere with these sorts of questions we cannot avoid venturing out across the philosophically and epistemologically boggy terrain of what is meant by causality.

Causation

Interestingly, just a few years before Farr's 1861 letter, John Snow was using statistics – administrative statistics of deaths from cholera and the number of houses served by different water companies, combined with a house-to-house survey to establish which houses received water from which company – to show that cholera was a water-borne infectious disease. In other words, cholera was caused by people drinking dirty water and this causal conclusion came, at least in part, from a careful analysis of statistical data. And, of course, nearly a century later, Richard Doll and Bradford Hill (both later knighted), in perhaps the most famous example of the application of statistical methods to observational data to generate causal conclusions, accumulated overwhelming evidence to show that smoking tobacco is a cause of lung cancer.

At this point, some of you might be wondering why a social statistician working at the Institute of Education is peppering his professorial lecture with examples from medical statistics and epidemiology. There are, I think, three reasons for doing so. The first is that many of the methodological issues faced by social statisticians (and econometricians too) are similar to those that are familiar to epidemiologists. Hence, those of us working with social science data can often do a better job by having at least a passing acquaintance with developments in epidemiological methods. In particular, both groups spend much of their time analysing observational data rather than the experimental data generated by randomised controlled trials. The second reason is that the

route to establishing a valid causal link between an explanatory or risk variable and an outcome can be long and tortuous. Doll and Hill's conclusions were not immediately accepted by scientists; indeed, they were challenged by no less a figure than Sir Ronald Fisher because they were not experimentally based. But a series of studies with different designs all pointed to the same conclusion so that, eventually, the link between smoking and lung cancer was accepted even by some tobacco companies. Social scientists do not always appreciate the importance of replication. Finally, and perhaps most importantly, the work on the causes of both cholera and lung cancer was informed by theory in that it was possible to specify the mechanisms by which cholera was passed from water to humans, and the chemical processes involved in burning tobacco. No matter how sophisticated the statistical techniques used to analyse relations between variables, the translation of an association into a causal link can only be accepted if the link is theoretically grounded.

Was William Farr being too censorious of attempts to derive causal conclusions from observational data? We have learnt from epidemiological investigations that a combination of sound theory, careful research design and detailed statistical analysis can generate conclusions that most reasonable people are prepared to accept are causal. And there is no doubt that social scientists ask many causal questions, arguably more than their counterparts in the physical sciences do: does parental separation lead to later behavioural problems for their children; does putting more police on the beat reduce street crime; do pupils make more academic progress if they are taught in smaller classes? But, as Freedman (1999) points out, answers are not guaranteed and there are many examples of apparently causal links turning out to be no more than associations. Moreover, social scientists – especially those working with educational data – cannot always draw on strong theory to guide them in their search for causal conclusions.

Let me now turn to what I mean by 'cause'. I do so with some trepidation, recognising that causality is a topic that has received abundant attention from philosophers, scientists and statisticians alike. My own views have been shaped by a number of authors but particularly by Cook and Campbell (1979), and by Sir David Cox (1992) and Cox and Wermuth (2001).

Those of us who analyse quantitative social science data have a working definition of causality that runs more or less along the following lines. A relation between two variables is deemed to be causal if:

(a) it is theoretically plausible;
(b) if a change in the outcome (or response or effect) of interest (y) is more likely following a change in causal variable (x) [as experienced by cases in class P] than if there is no change in x [as experienced by cases in class Q];
(c) the changes in (b) hold even after controlling (i) for other changes that are contemporaneous with the change in x and (ii) for fixed characteristics that are associated with the change in x, variables that are often known as confounders that can create spurious causal inferences.

Thus, taking up smoking will increase your chances of contracting lung cancer later in life, but not all smokers get lung cancer and some non-smokers still get the disease. In other words, causal relations in the social sciences are probabilistic just as statistical predictions are. Moreover, there can be more than one cause of the same outcome: for example, child A's performance at school improved because his parents' socio-economic circumstances improved, whereas child B started to do better because her classmates stopped bullying her. It is widely accepted that we cannot treat time or earlier (or lagged) measures of the effect (i.e. $y_{t-k}, k \geq 1$) as causal variables. The emphasis placed on an earlier change in x leading to a later change in y also rules out ascribed characteristics, notably sex, age and ethnic group, as causes. This position is, however, open to criticism especially when we think about discrimination in, for example, the labour market. There is evidence that women are paid less than men just because they are women, and that older people and black people are denied job opportunities because of their age or colour. I return briefly to this difficulty with the way causality is defined in the next section.

Causal relations are established more easily by conducting randomised experiments than they are by analysing data from observational studies, but the

opportunities for randomisation, although occurring more frequently than is often supposed, remain somewhat limited. Instead, *faute de mieux*, we try to do in the analysis something that is, in principle, more easily done as part of the design of a study and our working definition of causality reflects this. It is brought out by the implied counterfactuals of what would have happened to cases in class P if x had not changed and what would have happened to cases in class Q if x had changed. This is one of the reasons for the perceived strength of longitudinal data: repeatedly measuring the same cases over time makes it possible both to relate a change (or no change) in the cause (x) to a change in the effect (y), and to eliminate the possibly confounding influences of other variables.

We might suppose that the use of longitudinal data to generate causal inferences from observational studies in the social sciences is a relatively modern activity. In fact we can go back to the very end of the nineteenth century to find what is surely the first example of this kind of analysis. In 1899, G. Udny Yule, then only aged 28, presented a paper to the Royal Statistical Society on the causes of changes in pauperism (Yule, 1899). Yule was certainly aware of the difficulties of answering questions of this kind because, in 1897, he had written:

> The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics.
>
> (Yule, 1897: 812)

It is a humbling experience to read Yule's 1899 paper, and the discussion that followed it which, in the best traditions of the Royal Statistical Society, contains many insightful contributions. His data came from administrative sources: the

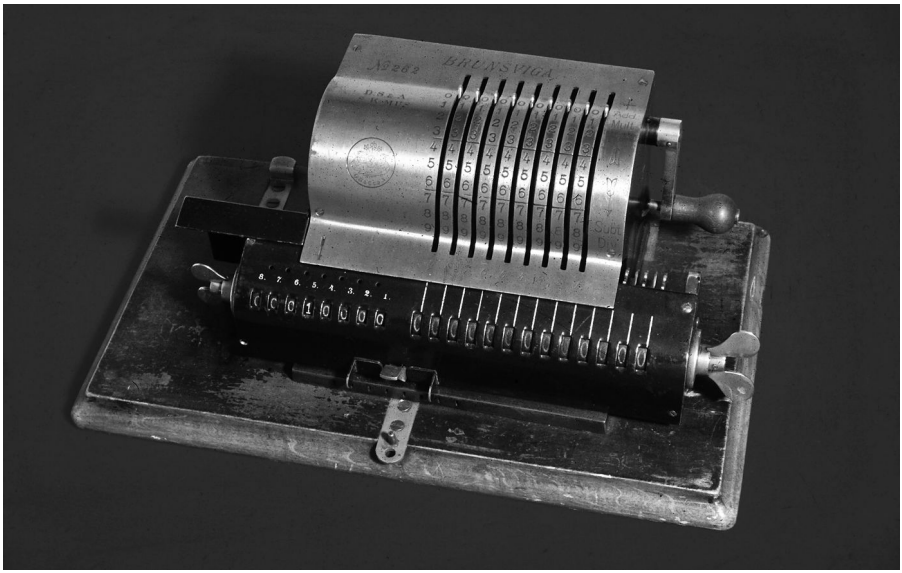
580 or so poor law unions in England between 1871 and 1891. These poor law unions were geographical areas not so different from local authorities today and they also functioned as registration districts. They administered relief in two ways: indoor relief (workhouses) and outdoor relief (benefits in money and kind). Yule's underlying hypothesis was that outdoor relief – which the Poor Law Amendment Act of 1834 had in fact tried to eliminate – created pauperism, and so he expected to find a reduction in pauperism in those unions that had reduced the proportion of outdoor relief. His data were longitudinal in that he collected data for the same unions on three occasions, the population census years of 1871, 1881 and 1891. He related changes in the proportions of outdoor relief to changes in pauperism rates and he controlled for (a) changes in the population (a proxy for prosperity) and (b) changes in the proportion of old people (who were more likely to be paupers). His underlying model was an easily recognisable multiple regression model:

$$y_i = \sum_{k=0}^{3} b_k x_{ki} + e_i$$

where the effect is the percentage ratio change in pauperism for union i, the cause is the corresponding change in 'out-relief' ratio (x_1), and the controls are changes in the population (x_2) and the proportion of the elderly (x_3). Yule's particular interest was in the estimate of b_1 – the coefficient for change in 'out-relief' ratio – which he interpreted, albeit cautiously, as the causal effect of changes in poor law administration on rates of pauperism and which he expected to be positive: as the cause goes up (or down) so the effect goes up (or down).

Yule fitted the model by the method of least squares, using a mechanical arithmometer, to eight datasets: separately for the periods 1871–81 and 1881–91 for unions in four kinds of areas. His expectations were confirmed in that the estimates of the coefficients of interest, b_1, were generally large and positive and he concluded:

> It seems impossible to attribute the greater part, at all events, of the observed correlation between changes in pauperism and changes in



Brunsviga calculating machine, 1892 © The Science Museum

> out-relief ratio to anything but a direct influence of change of policy on change of pauperism, the change in policy not being due to any external causes such as growth of population or economic changes.
>
> (Yule, 1899: 277)

It is, I think, a remarkable paper not least because Yule assumed, without any discussion, that his analysis required longitudinal data and the corresponding analyses of *changes* in the effect related to *changes* in the putative cause. Anyone starting out on a career in social statistics would benefit from studying the article; it is a reminder both of how little but also how far we have travelled. We are still using essentially the same models that Yule used over 100 years ago, although we can now estimate them in seconds rather than in days: the multiple regression workhorse has stood the test of time. On the other hand, as statisticians, we would blanch at Yule's willingness to interpret the sizes of regression coefficients that have such large standard errors; at his over-interpretation of the intercept term (b_0); we would want to consider different

functional forms and different link functions for the regressions; we would wonder about the residuals from the fitted model; and we would be concerned by a possible ecological fallacy – whether the effects on *individuals* of taking away outdoor relief are the same as the effects at the poor law union level. And as researchers, we would ask ourselves about the distinction between being a pauper and being poor; and whether the causal effect might perhaps be in the other direction: out-relief declined because pauperism was declining.

Let us now move from the end of the nineteenth century to the beginning of the twenty-first and to another research question about poverty, one that encapsulates many of the difficulties of establishing causal relations from longitudinal observational data. We know that there is an association between household income and school attainment: children from poor families do less well at school. We might expect that children who experience an improvement in their parents' economic circumstances will then start to do better at school, and those who experience a decline will make less progress. We can test this causal proposition and, with funding from the Department for Work and Pensions (DWP) and drawing on data from the cohort studies, this is something that I, with CLS colleagues Denise Hawkes and Constantinos Kallis, have been considering over the last three years.

The essential problem we face is that to estimate the effect of a change in income on educational outcomes we need to observe a range of income changes, both positive and negative and including no change. But many of the income changes that we observe are, in one way or another, bound up with the characteristics of the sample members. For example, one person's income goes up as a result of choosing to undertake training whereas another person's goes down because they were unable or unwilling to update their skills. These are endogeneous changes; the people select themselves into situations that affect their worth in the labour market and these people are likely to be more or less positive about the value of education for their children. We would much prefer to look at the effects of exogeneous changes but, unfortunately, our samples contain, for example, very few winners of lottery prizes. Longitudinal data are essential if we are to eliminate the effects of self-selection but they are not a panacea. We have to think very carefully about which variables we measure,

how we specify the underlying model and which statistical techniques we use to estimate it.

The model we have been working with is shown in Figure 2. We assume that the effects of fixed unmeasured or unobserved variables act through income just at time t and not also through income at time t+1. These are variables like 'motivation' that are often bundled together and labelled 'unobserved heterogeneity'; their potential for biasing estimates of causal effects if the above assumption does not hold exercises many researchers and has led to a variety of suggestions about how to avoid their perils. Economists, in particular, look to replace the putative cause by one or more instrumental variables – variables that are correlated with the cause but have no direct relation to the effect. I am frequently delighted by the imagination shown by my economist colleagues as they exploit policy changes, even political upheavals, in their search for valid instruments. Yet I remain sceptical about the value of their efforts; it seems to me that the better the instrument in terms of its separation from the effect, the lower the correlation with the cause of interest, and hence the common finding that standard errors attached to estimates generated from instrumental variable estimation are very high.

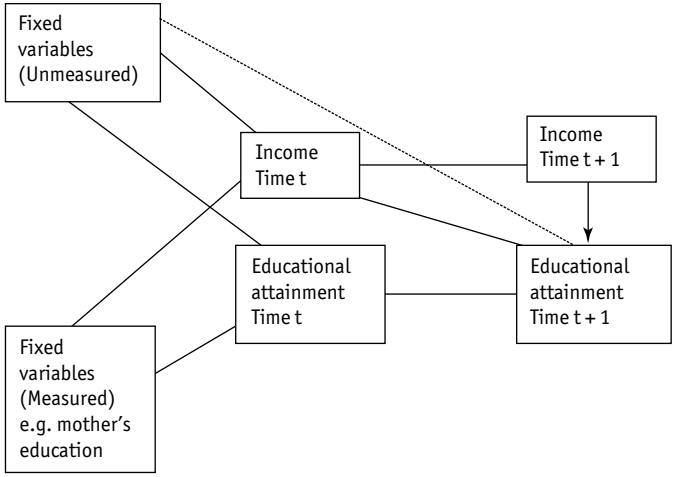


Figure 2 Model for relating changes in income to educational progress

Another lively area of debate is over the use of fixed and random effects models. Suppose we are in the fortunate position of having multiple repeated measures of our putative causes, our effects and a set of potentially confounding variables. (In our work, we have had to make do for the most part with just two measures.) We can write the model as:

$$y_{ti} = b_{0i} + b_1 x_{ti} + \sum_{j=1}^{J} c_j z_{ji} + \sum_{k=1}^{K} d_k z^*_{kti} + e_{ti},\ t = 1..t_i;\ i = 1..n$$

where t are the measurement occasions and n is the sample size.

Our interest focuses on b_1 as this represents the causal effect of x on y if our model is properly specified. The z variables are fixed characteristics, measured and unmeasured, of the sample members, correlated with x and believed to influence y that we must control for, the z^* are those time-varying variables that can be treated as confounding variables (see below). It is what we should do with the b_{0i} that generates the most discussion. One approach is to treat them as a set of fixed effects or dummy variables, one for each member of the sample. By doing so, we eliminate the problem of unmeasured confounding variables that are fixed over time but we pay a price for so doing: for large samples, we have to estimate a lot of nuisance parameters and we lose any possibility of estimating interactions between the cause and the fixed confounders – so-called moderator effects – because the z variables are subsumed within the individual effects. A variant of the fixed effects approach is to model the deviations of each y_{ti} from its individual mean y_i.

Alternatively, we can treat the b_{0i} as random effects so that we just estimate their variance within what is now a two-level model (occasions within individuals), and that model allows us to generalise to a population. We can extend it to allow the causal effect b_1 to vary across individuals, perhaps systematically with the fixed characteristics z. We can also include higher levels such as schools or neighbourhoods: might, for example, the effect of a change in income be stronger in some neighbourhoods than in others? The random effects model is both more parsimonious and more flexible than the fixed effects model, but the price of these advantages is the possibility that unmeasured confounders will lead to a biased estimate of the causal effect. There is, of course, a trade-off

between bias and statistical precision and methodological comparisons between the two approaches on the same dataset would seem to be warranted. For one example, see Plewis (2001).

At this point, I want to raise some questions about unmeasured confounders. Are they really fixed within individuals over time? Are they variables we know about but were not able to measure? This is certainly a problem for anyone carrying out secondary data analysis. Or are they variables that we think might be important but are not sure how to measure – these rather nebulous concepts like 'motivation' and 'expectations'? Now, it seems to me that, rather than brushing these variables under the fixed-effects carpet, we should be developing measures of them and including them in our models just as we would more straightforward confounders like levels of education.

We do also have to be very careful about which time-varying variables we include in our model. If we think about changes in income related to educational progress then we know that some of those changes will arise because of changes in family structure, in particular the change from a two-parent to a one-parent family or vice versa. Do those changes mediate or explain the causal relation or do they confound it? The same kinds of questions arise when we look at differences between ethnic groups in attainment and progress. Should we control for differences between the groups in socio-economic circumstances, or should we not recognise that these differences might have arisen because of discrimination and racism and are therefore endogeneous? There are no straightforward answers to these questions but it is important to recognise that it is possible to 'over control' and thus to eliminate causal effects that are important. An awareness of the broader social and political climate within which our research and our statistical modelling takes place is, I think, essential and can assist us to make difficult decisions about model specification.

It is, of course, absolutely correct that causal models generated from observational data are subject to close scrutiny. Like all models and theories, they need to be replicated on different samples and for different populations; they will always be tentative and should be falsifiable. Much of the scrutiny comes from the search for omitted confounders that are, in turn, generated by self-selection mechanisms. This search should however, as Bross (1960) pointed out, be

guided by a sense of statistical responsibility. Commenting on Fisher's rejection of a causal link between cigarette smoking and illness, he wrote:

> Instead of attempting to make the self-selection hypothesis tenable, Fisher simply dismissed the entire body of epidemiological data. . . . He did so on the basis that the data do not meet certain theoretical standards for 'properly controlled experimentation'. This seems to me a gross violation of the empirical spirit of modern science *and* of modern statistics. It raises the theory of statistics, e.g. randomization, to the level of dogma.
>
> (Bross 1960: 396)

In other words, we should not reject a finding just by documenting some omitted variables; we should also have a tenable counter-hypothesis that can explain the observed association, a point that some of us were at pains to point out in a Radical Statistics pamphlet published 25 years ago (Radical Statistics Education Group, 1982).

Explanation

As my colleague Ros Levačić pointed out in her 2004 Professorial Lecture (Levačić, 2005), it is very useful to establish that x is a cause of y but that is usually only part of the full story. Ideally, we would like to know what the underlying processes are that generate this causal effect, what Goldthorpe (2001) describes as 'causation as generative process' and what are often referred to as mediating variables. I return to two of my examples to illustrate what the search for explanations might involve.

Considering first the relation between changes in socio-economic circumstances at home and educational progress, we can entertain at least two theoretically plausible explanations for any causal effect that we might find. The first is that some of an observed increase in disposable income can be spent on goods and services that might be expected to have an effect on educational

progress: books and educational games, extra tuition, visits to museums, etc. The second is that a decline in parents' economic position, as often happens after a separation, is associated with the amount of time that parents can give to their children's education in the form of help and guidance at home. If we find evidence for the first explanation then we might say that changes in household income operate through changes in patterns of consumption to produce changes in educational progress (Figure 3) – consumption is a mediating effect. If the effect appears to operate through parental time allocations then we might want to use that as an argument for treating changes in family structure as a confounding variable (because these probably change the amount of time available) or we might want to argue that changes in income and changes in family structure are both causes. We might then fit separate models for each of them.

Let us now turn to the current educational fact that in primary school boys make less progress in English than girls, and this is particularly so for boys from an African-Caribbean background. This fact requires an explanation, although now we want to explain an association that, because it relates fixed ascribed variables to an outcome, might not be regarded as causal in the sense discussed

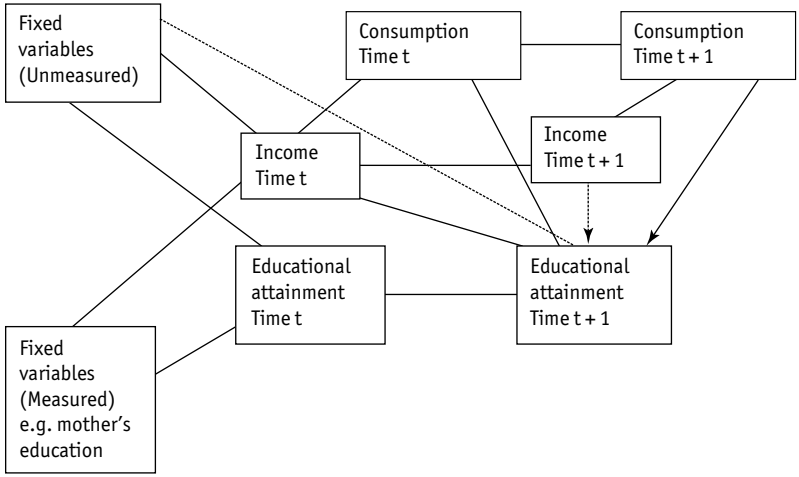


Figure 3 Model for relating changes in income to educational progress via changes in consumption

earlier. If we were able to establish that those variables that are causally related to educational progress are unequally distributed between the ethnic and gender groups of interest then we would have strong grounds for claiming an explanation in the sense of being able to explain away the association. This was essentially the approach adopted in the Infant School Project, but we were not able to find any variables – neither those describing what went on in the classroom nor those describing what went on at home – that separated the two groups. But, as I pointed out earlier, we were handicapped by the relatively small sample. Although this handicap does not apply to the National Pupil Database, the NPD is not a suitable vehicle for establishing explanations because it contains no information about pupils' educational experiences at school or at home. Consequently, explanations for differential progress between the sexes and ethnic groups remain somewhat elusive.

In both these examples, we first need to generate evidence that x (or $x_1 \ldots x_k$) causes y. In the first example, our mediating variables explain the causal effect in the sense that its magnitude might be reduced by the inclusion of the mediating variables in a more complex model (i.e. Figure 3). However, in the second example, it is the unequal distribution of the putative cause or causes across groups defined by ascribed characteristics like sex and ethnic group that provide the explanation, in that the association with these ascribed characteristics would then be eliminated.

If we are in the fortunate position of having at our disposal a dataset that includes all of the relevant variables – the effects, the causes, the confounders, the mediators and the moderators – then we can model them. But rarely are we so lucky. It can be difficult to measure the potential mediating variables in quantitative studies like the cohort studies with their large samples but with restrictions on how much data can be collected. These restrictions come from financial constraints and, importantly, from the need not to over-burden the respondents. Instead, we might do better to incorporate some smaller-scale but more detailed studies into our large longitudinal studies in order to improve our understanding of underlying processes. Questions about intra-household transfers, patterns of consumption, use of time, and family members' reports of changes in behaviour in response to changes in income all require rather

detailed questions and complicated research instruments. Quantitative social science is sometimes criticised for not being able to provide insights into processes. I think the criticism is unwarranted, as I have demonstrated here. Nevertheless, combining large-scale investigations based on probability samples of the population with more detailed but smaller-scale studies, perhaps based on theoretical sampling, offers the potential for more complete analyses in circumstances of this kind.

Conclusions

Quantitative researchers in the social sciences have come to appreciate the strengths of longitudinal data. No longer do the problems I was addressing in my SSRC Fellowship loom so large. On the other hand, there are still plenty of outstanding issues for statisticians and methodologists to get their teeth into. To list just a few:

- What are the best ways of using information from statistical prediction rules to allocate resources?
- How can we best describe differences *between* cohorts?
- How long does it take for a cause to be translated into an effect, and how do we represent that lag in causal models that are based on data generated by designs with predetermined intervals between waves?
- To what extent might causal effects found at one level (the pupil say) be mediated by changes at a higher level such as the school?

Just as there are many outstanding questions about longitudinal analysis, so there are questions about design and data collection, but I have not been able to consider these important issues in this lecture.

We need more highly-trained social statisticians and quantitative researchers to tackle questions like these. We need them to produce reports and papers that might be on the dry side but which should certainly not be desiccated: writing

that engages readers in the way that Florence Nightingale set out to achieve despite William Farr's restraining hand on her shoulder.

Acknowledgements

I would like to thank Jane Elliott, Harvey Goldstein and Heather Joshi for helpful comments on earlier versions of this lecture, and also the many researchers with whom I have collaborated during my career at the Institute and who have helped to shape my ideas.

References

Bross, I.D.J. (1960) 'Statistical criticism'. *Cancer*, 13, 394–400.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cox, D.R. (1992) 'Causality: some statistical aspects'. *Journal of the Royal Statistical Society A*, 155, 291–301.

Cox, D.R. and Wermuth, N. (2001) 'Some statistical aspects of causality'. *European Sociological Review*, 17, 65–74.

Diamond, M. and Stone, M. (1981) 'Nightingale on Quetelet'. *Journal of the Royal Statistical Society A*, 144, 66–79.

Edwards, A., Barnes, M., Plewis, I. and Morris, K. *et al.* (2006) *Working to Prevent the Social Exclusion of Children and Young People*. Department for Education and Skills Research Report No. 734.

Economic and Social Research Council (2006) *SSRC/ESRC: The first forty years*. Swindon: ESRC.

Freedman, D. (1999) 'From association to causation: some remarks on the history of statistics'. *Statistical Science*, 14, 243–58.

Goldthorpe, J.H. (2001) 'Causation, statistics and sociology'. *European Sociological Review*, 17, 1–20.

Hawkes, D. and Plewis, I. (2006) 'Modelling non-response in the National Child Development Study'. *Journal of the Royal Statistical Society A*, 169, 479–91.

Kuh, D., Head, J., Hardy, R. and Wadsworth, M. (1997) 'The influence of education and family background on women's earnings in midlife: evidence from a British national birth cohort study'. *British Journal of Sociology of Education,* 18, 385–405.

Levačić, R. (2005) *The Resourcing Puzzle*. Professorial Lecture. London: Institute of Education, University of London.

Plewis, I. (1988) 'Assessing and understanding the educational progress of children from different ethnic groups'. *Journal of the Royal Statistical Society, Series A*, 151, 316–26.

—— (1991) 'Pupils' progress in reading and mathematics during primary school: associations with ethnic group and sex'. *Educational Research*, 33, 133–40.

—— (1996) 'Statistical methods for understanding cognitive growth: a review, a synthesis and an application'. *British Journal of Mathematical and Statistical Psychology*, 49, 25–42.

—— (2001) 'Explanatory models for relating growth processes'. *Multivariate Behavioral Research*, 36, 207–26.

—— (in preparation) 'The role of risk and protective variables in the construction and use of statistical prediction rules'.

Radical Statistics Education Group (1982) *Reading Between the Numbers*. London: British Society for Social Responsibility in Science Publications.

Swets, J.A., Dawes, R.M. and Monahan, J. (2000) 'Psychological science can improve diagnostic decisions'. *Psychological Science in the Public Interest*, 1, 1–26.

Tizard, B., Blatchford, P., Burke, J., Farquhar, C. and Plewis, I. (1988) *Young Children at School in the Inner City*. Hove: Lawrence Erlbaum.

Yule, G.U. (1897) 'On the theory of correlation'. *Journal of the Royal Statistical Society*, 60, 812–54.

—— (1899) 'An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades (Part 1) (with discussion)'. *Journal of the Royal Statistical Society*, 62, 249–95.