SECTION 7
QUANTITATIVE RESEARCH

Prepared for the MA Boardby Roger Gomm

11 INTRODUCTION

The term 'quantitative research' is subject to different definitions. For the purposes of this section we shall proceed as if it referred to:

1 The search for causal relationships conceptualized in terms of the interaction of 'variables', some of which (independent variables) are seen as the cause of other (dependent) variables.

2 The design and use of standardized research instruments (tests, attitude scales, questionnaires, observation schedules) to collect numerical data.

3 The manipulation of data using statistical techniques.

Our main emphasis in this section will be on the forms of analysis used by quantitative researchers, rather than on the data collection techniques they employ. This is because the latter are constrained to a considerable degree by the requirements of quantitative analysis. Most obviously, the data need to be in numerical form - measurements of the intensity and/or frequency of various phenomena. Some of this type of data is readily available in the form of published or unpublished statistics: for example, school examination results, figures for absenteeism, etc. Often, though, researchers have to produce the data themselves. This may be done by means of laboratory or field experiments, or through the use of surveys involving structured questionnaires or systematic observation schedules.

Phenomena vary, of course, according to how easily and accurately they can be measured. It is one thing to document the number of A levels obtained by each member of the sixth form in a school in a particular year. It is quite another to document the proportion of sixth formers from working-class and from middle-class homes. There are troublesome conceptual issues involved in identifying membership of social classes, as we shall see; and collecting accurate information on which to base assignment to social classes is much more difficult than finding out how many A levels were obtained. Although we shall concentrate primarily on techniques of analysis, the threats to validity involved in the process of collecting data must not be ignored. We shall have occasion to discuss these at several places in this section.

In the early parts of the section we shall make extensive use of a relatively simple example of quantitative work drawn from Stephen Ball's *Beachside Comprehensive,* a book whose main approach is qualitative rather than quantitative (Ball, 1981). In this study Ball is primarily concerned with finding out whether, or how far, the principles of comprehensive education have been implemented at a particular school, to which he gives the pseudonym 'Beachside'. One of the main criticisms of the selective educational system in Britain of the 1950s and the early 1960s was that it disadvantaged working-class children because selection for different kinds of education occurred so early. The move to comprehensive schools was intended, in part, to overcome this problem. In carrying out this research in the late 1970s, Ball was particularly interested in how far it was the case that working-class and middle-class children had an equal chance of educational success at schools like Beachside Comprehensive. We shall look at only a very small part of his work, where he examines whether the distribution of working-class and middle-class children to bands on their entry to the school shows signs of inequality.

Activity 43 (allow 1 hour)

You should now read 'Banding and social class at Beachside Comprehensive' by Stephen Ball (Article 7 in the Offprints Reader). As you do so, make a list of the main claims he puts forward and of the types of evidence he presents. You don't need to go into much detail at this point. Don't worry about statistical terminology that you don't understand; your aim should be simply to get a general sense of the structure of Ball's argument. However, this will take careful reading.

In the extract Ball argues that there is evidence of bias against working-class pupils in their allocation to bands at Beachside. He supports this by comparing the distribution of middle-class and working-class children across Bands 1 and *2* with their scores on the NFER tests of reading comprehension and mathematics. We shall be considering this evidence in some detail later, but first we need to give some attention to the nature of the claim he is making.

7.2 CAUSAL THINKING

Very often researchers are looking for the causes of phenomena, and those who use quantitative methods tend to do so in a characteristic way.

Activity 44 (allow 3 hours)

You should now read 'How to think about causality' by J. Hage and B. F. Meeker (Article 6 in Reader 1). This is quite a difficult article, but it is very useful in the way it maps out the various sorts of causal processes to be found in the social world.

When you have finished reading it, construct a causal diagram along the lines of those provided by Hage and Meeker to represent Ball's argument about the relationship between social class and allocation to bands at Beachside Comprehensive. Do this before you read on.

Hage and Meeker's thinking might be applied to the section of Ball's work that you read in the form of the causal-network diagram shown in Figure 4.

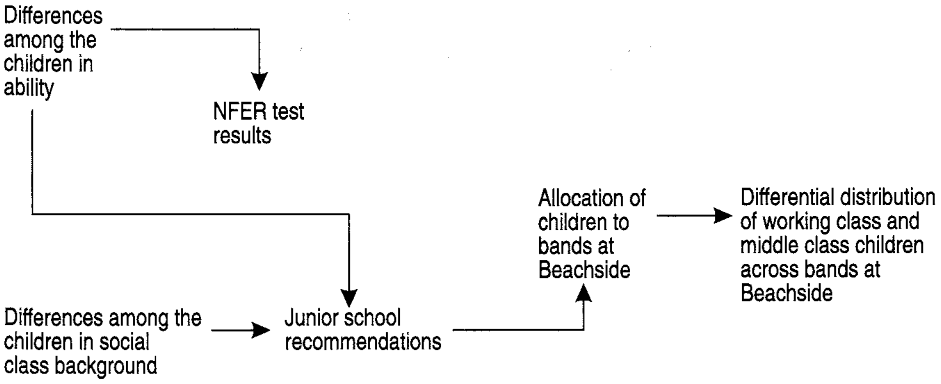

Differences among the children in ability
NFER test results
Allocation of children to bands at Beachside
Differential distribution of working class and middle class children across bands at Beachside
Differences among the children in social class background
Junior school recommendations

Figure 4 A causal-network diagram of Stephen Ball's argument

Your diagram may not be exactly the same as ours, since there is some scope for variation, but we believe that this causal model captures the main lines of Ball's argument.

It is worth noting that the diagram presupposes a temporal sequence in that all the arrows go one way, from left to right. The one solid fact about causality is that something can only be caused by something that precedes it. Thus, the left-hand side of the diagram must refer to earlier events and the right-hand side to later events. In our case 'social class' must mean something about pupils' home background which existed prior to their allocation to bands. Only in this sense can social class be said to have a causal relationship to allocation to bands. Of course, in principle at least, the causal network could be extended almost indefinitely backwards or forwards in time.

In quantitative analysis the terms 'dependent' and 'independent' variables are frequently used to conceptualize causal relationships. Variation in the independent variable represents the cause, while variation in the dependent variable is the effect we are interested in explaining. In Figure 4, pupils' social class and ability are the independent variables, and their position in the bands represents the dependent variable.

The 'time rule' for causal analysis means that changes in independent variables must occur before the changes in the dependent variables they are held to cause. Beyond this, the way the terms are used relates to the explanatory task at hand. Thus we could move back a step and see differences in pupils' ability as the dependent variable, in which case differences in social background (and other things) would be the independent variable. On the other hand, we could treat social-class differences in upbringing as the dependent variable and look for factors that explain this, treating these as independent variables. The reference of the terms 'independent' and 'dependent' variables, then, derives from the explanatory task being pursued.

Another point to notice is that we could go on adding to our independent variables, perhaps for ever. In setting up an explanatory model, a causal network, we are selective about what we include among the independent variables. Thus, Ball was concerned with the extent to which the allocation of pupils to bands at Beachside was affected independently by social class, rather than being based solely on pupils' ability. He was looking to see if middle-class children were more likely than working-class children to be allocated to Band 1 over and above any differences in ability. This research problem determined his selection of independent variables: social-class background and pupils' ability. (As you will realize, there is a close parallel between Ball's work here and that of Troyna, which we examined in Part 1, Section 3.3.)

In our diagram, we have also included the junior schools' recommendations as a factor mediating the effects of those two other variables. The teachers from Beachside who carried out the initial allocation to bands had no knowledge of the children other than what was available to them via the junior schools' records and recommendations. If they had had such knowledge then we should have needed to include direct causal relationships both between ability and band allocation and between social class and band allocation.

It is worth thinking a little about the nature of Ball's hypothesis. As we indicated earlier, his concern is with the extent to which there is inequality (or, more strictly speaking, inequity) in allocation to bands in terms of social class. This is quite a complex issue, not least because it relies on assumptions about what is believed to be equitable and, more generally, about what is a reasonable basis for the allocation of pupils to bands in a school. It suggests that what is partly at issue in Ball's argument is a question about the basis on which band allocation *should be* made, not just about how it is actually carried out. In our discussion here, however, we shall treat Ball's hypothesis as factual; as concerned with whether or not social class has an effect on allocation to bands over and above the effects of differences in ability.

Operationalization

In order to test his hypothesis, of course, Ball had to find measures for his variables. Thus, in the extract you read, he is not directly comparing the effect on band allocation of the social class and the ability of the children. Rather, he compares the band allocation of children who obtained scores in the same range on the NFER tests for reading and mathematics, whose fathers are manual and non-manual workers. He has 'operationalized' social class in terms of the difference between households whose head has a manual as opposed to a non-manual occupation and has 'operationalized' ability in terms of scores on NFER achievement tests. This, obviously, raises questions about whether the variables have been measured accurately.

How well does the occupation of the head of household (categorized as manual or non-manual) measure social class?

Activity 45 (allow 5 minutes)

What potential problems can you see with this operationalization?

This is by no means an uncontroversial question. There are several problems. For one thing, there are conceptual issues surrounding what we mean by social class. For instance, social class differences can be viewed as a dichotomy between opposed groups with different relationships to the means of production (as in the writings of Karl Marx) or in terms of a scale representing variation in some set of features (level of income, social status, etc.). Furthermore, Ball's operationalization seems to involve allocation of children to social classes on the basis of their fathers' occupations. Yet, of course, many mothers have paid employment and this may have a substantial impact on households. There has been much debate about this (see Marshall *et al.*, 1988 for a discussion). Another question that needs to be asked is whether the distinction between non-manual and manual occupations accurately captures the difference between working class and middle class. Finally we might raise questions about the accuracy of the occupational information that was supplied by the pupils and on which Ball relied.

Activity 46 (allow 5 mintues)

Can you see any likely threats to validity in Ball's operationalization of ability? Note down any that you can think of before you read on.

To answer this question, we need to think about what the term 'ability' means in this context. It is worth looking at how Ball introduces it. He appeals to the work of Ford, arguing that controlling for measured intelligence is the most obvious way of testing for the presence of equality of opportunity. If the impact of social class on educational attainment is greater than can be explained by the co-variation of social class and IQ, then the existence of equality of opportunity must be called into question, he implies. Effectively, Ball is asking whether the allocations to bands were fair, as between children from different social classes, and takes a 'fair allocation' to be one that reflects differences in intelligence.

One point that must be made in the light of this is that Ball's operationalization of ability relies on achievement tests in reading and mathematics, not on the results of intelligence tests (these were not available). We must consider what the implications of this are for the validity of the operationalization.

A second point is that to the extent that the allocations were based on predictions about likely academic success, the teachers making the allocations would have been unlikely to rely solely on the results of achievement tests. They would have used those scores, but also any other information that was available, such as their own and other teachers' experience with the children. Furthermore, they were probably interested not just in intelligence and achievement but also in the level of motivation of the children, since that too seems likely to have an effect on future academic success. In fact, you may remember that Ball notes that Beachside School allocated children to bands on the basis of the primary school's recommendations and that 'test scores were not the sole basis upon which recommendations were made. Teachers' reports were also taken into account' (Offprints Reader, p. 70).

What this indicates is that the primary schools did not regard test scores as in themselves sufficient basis for judgments about the band into which pupils should be placed. Given this, it would not be too surprising if Ball were to find some discrepancy between band allocation and test score, although this discrepancy would not necessarily be related to social class.

Over and above these conceptual issues, all tests involve potential error. Ball himself notes the problem, commenting that 'to some extent at least, findings concerning the relationships between test-performance and social class must be regarded as an artefact of the nature of the tests employed' (Offprints Reader, p. 72). These problems are certainly not grounds for rejecting the data, but they should make us cautious in handling them.

It is important to notice that these sorts of problems arise with all operationalizations, to one degree or another. We have to make judgments about whether the level of likely error is such as to undermine the validity of any conclusions drawn on the basis of the particular operationalizations employed. This is a matter about which researchers may disagree.

Activity 47 (allow 3 hours)

You should now read 'Questioning ORACLE': an assessment of ORACLE'S analysis of teachers' questions' by J. Scarth and M. Hammersley (Article 11 in Reader 2). As you do so, make a list of their criticisms of the ORACLE research. Many of these relate to problems of operationalization.

There are, then, some serious questions to be raised about the operationalization of both social class and of ability in Ball's article. For the purposes of our discussion, however, let us assume that the measurements represented by Ball's figures are accurate and that ability (as he operationalizes it) is the appropriate criterion for band allocation.

7.3 CO-VARIATION

In quantitative research the evidence used to demonstrate a causal relationship is usually 'co-variation'. Things that vary together can usually be relied upon to be linked together in some network of relationships between cause and effect, although the relationships may not be simple or direct.

Activity 48 (allow 5 minutes)

Look again at Table 2.5 in the extract from Ball which you read earlier (Offprints Reader, p. 71). Does this table display co-variation between social class (in Ball's terms) and allocations to the two forms that he studied?

The answer is that it does. We can see this just by looking at the columns labelled 'Total non-manual' and 'Total manual'. Form 2CU contains 20 pupils from homes of non-manual workers and 12 from those of manual workers, whereas the corresponding figures for Form 2TA are 7 and 26. This pattern is also to be found if we look at the two top bands as a whole, rather than just these two forms (see Table 8).

Table 8 Distribution of social classes by ability bands

Source: Extracted from Ball (1981) Table N2, p. 293 (reproduced in Offprints Reader, p. 71)

Notice how we produced Table 8 by extracting just a small portion of the information that is available in Ball's Table N2. Doing so enables us to see patterns much more clearly. At the same time, of course, it involves losing quite a lot of other information, temporarily at least: for example, the differences between the social classes that make up the manual and non-manual categories. Having noted the general relationship, let us now include this extra information.

Table 9
Distribution of social classes across the second-year cohort at Beachside 1973-74

Source: Extracted from Ball (1981) Table N2, p. 293 (reproduced in Offprints Reader, p. 71)

Table 9 is much more difficult to read at a glance than Table 8. We need a strategy to make it easier to compare the various cells produced by the two co-ordinates, social class and band allocation. As we saw in Part 1, Section 3.3, there are ways of standardizing for group size, notably through the use of percentages. The formula for percentages is

$$\text{percentage} = \frac{f}{N} \times 100$$

where f = frequency in each cell and N equals, well what?

There are three possibilities in the case of Table 9. It could be the total of all pupils, or the total of pupils of a particular social class, or the total of pupils in a particular band.

Activity 49 (allow 15 minutes)

People so often misinterpret tables of percentages that it is worth doing this exercise to check your understanding. Think of each of the fragments (a), (b) and (c) below as parts of a new percentage table based on Table 9. Write a verbal description for each.

(a)		Social Class I	
	Band 1	3.7%	(*N* = total of pupils in all bands = 296)
(b)		Social Class I	
	Band 1	61.1%	(*N* = total of Social Class I pupils = 18)
(c)		Social Class I	
	Band 1	8.5%	(*N* = total in Band 1 = 129)

When you have done, this think about which of the three possibilities outlined above is most useful in relation to Ball's argument.

Here are our answers to Activity 49.

Item (a) can be described as the percentage of all pupils who are *both* classified as Social Class I *and* allocated to Band 1.

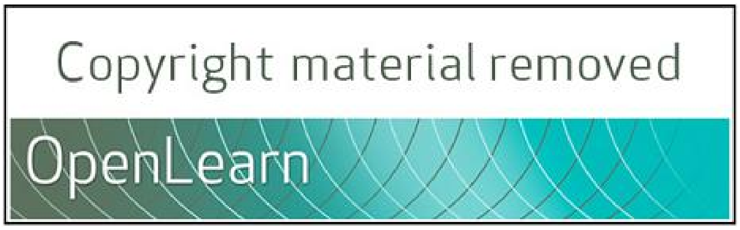
Item (b) is the percentage of all pupils classified as Social Class I who are allocated to Band 1.

Item (c) is the percentage of Band 1 places taken by Social Class I pupils.

For our current purposes (b) is the most useful percentage since it tells us something about the relative frequency with which pupils from a particular social class are allocated to a particular band. Item (c) might be interesting if you were concerned with the composition of bands, rather than with the chances of pupils gaining allocation to a band.

Table 10 is an expansion of item (b).

Table 10 Percentages of pupils from each social class allocated to particular bands



Source: Derived from Ball (1981) Table N2, p. 293 (reproduced in Offprints Reader, p. 71)

Now the data are converted into percentages we can, once again, see at a glance some co-variation between social class and band allocation. You could say, for example, that each pupil from Social Class I has six chances in ten of being allocated to Band 1, while each pupil in Social Class IV has under 1.5 chances in ten; the chances of a pupil from Social Class I being in Band 1 is four times that of the chances of a pupil from Social Class TV.

There are two problems with percentages, however. First, once you have converted numbers into percentages, there are strict limits to what you can do with them mathematically Percentages are mainly for display purposes. Secondly, once a number is converted into a percentage it is easy to forget the size of the original number. For example, look at the entries under Social Class V One hundred per cent of these pupils are in Band 2, but 100% is only five pupils. If just one of these five pupils had been allocated to Band 1, then there would have been 20% of pupils from Social Class V in Band 1 and only 80% in Band 2. Then, a higher percentage of pupils from Social Class V than from Social Class IV would have been in Band 1. Again, one extra pupil from Social Class I in Band 1 would raise their percentage to 66%. Where *Nis* small, small differences appear as dramatically large percentages. For this reason it is good practice in constructing tables of percentages to give the real totals (or base figures) to indicate what constitutes 100%. The misleading effects of converting small numbers to percentages can then be detected and anyone who is interested can recreate the original figures for themselves.

Activity 50 (allow 10 minutes)

The information provided in Tables 8 or 10 shows us that there is an association between social class (measured in the way that Ball measured it) and allocation to bands. What conclusions can you draw from this co-variation?

There are several ways in which this association could have been produced. First, it may be that teachers made allocations on the basis of criteria that favoured middle-class children. A second possibility is that the teachers allocated children on the basis of their ability, but that this is determined by, or co-varies strongly with, social class. There is also the third possibility that both social class and allocation to bands are caused by some other factor: in other words, that the association is spurious. We can illustrate these possibilities by use of causal-network diagrams (Figures 5 and 6).

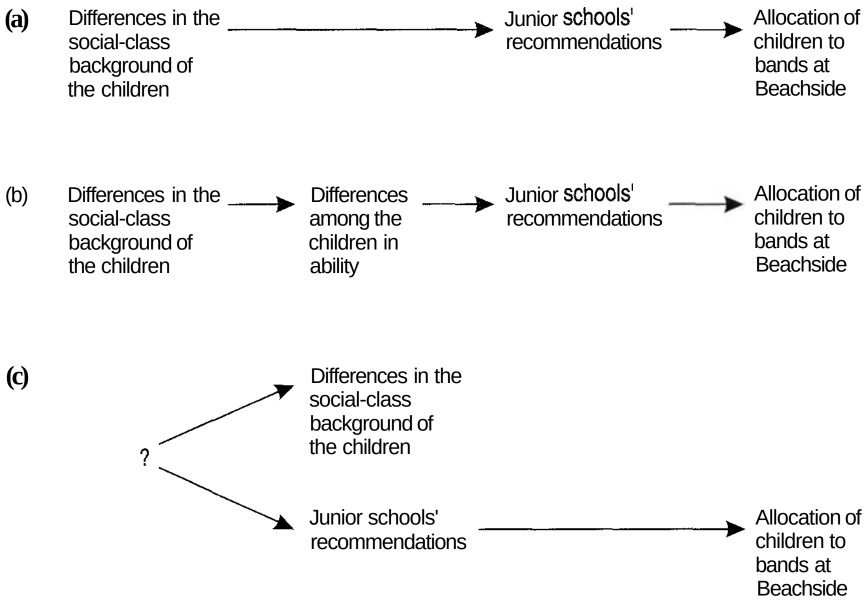

(a)
Differences in the social-class background of the children
Junior schools' recommendations
Allocation of children to bands at Beachside
(b)
Differences in the social-class background of the children
Differences among the children in ability
Junior schools' recommendations
Allocation of children to bands at Beachside
(c)
?
Differences in the social-class background of the children
Junior schools' recommendations
Allocation of children to bands at Beachside

Figure 5 Models of the relationship between social class and allocation to bands

It might be difficult to see what the mystery factor could be in Figure 5(c). It is worth noting, however, that some commentators have argued that ability is for the most part genetically determined and that ability to a large extent determines social class. On this view, we might get the causal network shown in Figure 6.

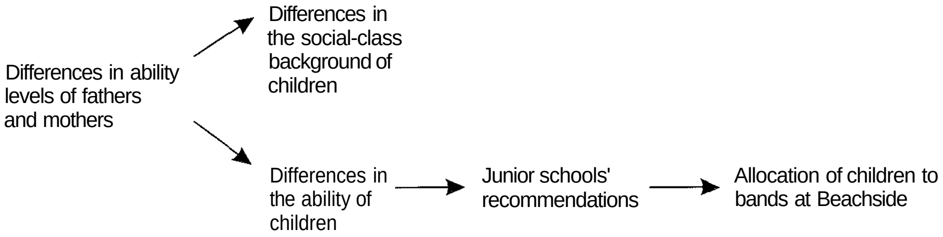

Differences in ability levels of fathers and mothers
Differences in the social-class background of children
Differences in the ability of children
Junior schools' recommendations
Allocation of children to bands at Beachside

Figure 6 A model assuming the inheritance of ability

A demonstration of co-variation between banding and social class, therefore, leaves room for different interpretations. This is generally true in social research. Here we are dealing with causal networks in which each item in the network may have a number of different possible relationships with the others. A crucial element in the art of planning a research study is to find ways of manipulating situations or data so that the independent effect of each possible causal factor becomes clear. The term used to refer to this kind of manoeuvre is 'controlling variables'.

7.4 CONTROLLING VARIABLES: EXPERIMENTATION

The scientific experiment is the classic example of a strategy for controlling variables. It is often said to involve physical control of variables, since it entails actual alteration of the independent variable of interest and the holding constant or minimizing of other factors likely to affect the dependent variable.

Suppose, as experimenters, we are interested in how teachers' judgements of pupils' ability and pupils' ability as measured independently of teachers' judgements each affect the position of pupils in some ability-banding system. Furthermore, we are interested in how far either or both create a pattern of distribution so that position in ability bands co-varies with social class, a pattern like that shown in Tables **8** or 10. As we have seen, social class may have **a** causative effect on allocation to ability bands through different routes because:

- Ability varies with social class and teachers recognize the differences in ability between pupils, which are then manifest in decisions about allocating pupils to bands.
- Teachers make assumptions about the ability of pupils of different social classes (which are not reflections of their ability) and implement these in decisions about allocating pupils to bands.

Our first problem in designing an experimental strategy is ethical. Even if people would agree to co-operate, it would be quite wrong to subject pupils to an experiment that was likely to have a real effect on their educational chances. We shall avoid this problem by conducting an experiment with fictional pupils who exist only on paper. The subjects for the experiment will be a group of teachers who have actually been involved in allocating pupils to bands. In the experiment we are going to ask them to make pencil and paper decisions about the allocation of fictional pupils to three ability bands: top, middle, and bottom. For the purpose of the experiment we shall divide the teachers into three groups, either at random or by matching. In other words, we shall try to ensure a similar balance in each group for age and gender, at least, and preferably also for other relevant variables, such as kind of subject taught, status in school, etc. Either way, the aim is to make each group of teachers as similar as possible so as to rule out the effects of their personal characteristics.

The two most important features of our fictional pupils will be their ability, as measured by some standardized test, and their social class, as designated, say, by the occupation of the head of their households. For a real experiment we might want a more elaborate system of classification, but for demonstration purposes let us say that the pupils will be designated as of high, middle or low ability and as coming from manual or non-manual backgrounds. We shall also make sure that there will be exactly the same number of pupils in each ability band from each social class.

As it stands, the object of our experiment is likely to be transparently obvious to the teachers involved, so we shall need to dissimulate a little. To do this we shall provide the pupils with genders, heights, weights and other details, but ensuring that the same characteristics always occur with the same frequency for each cell of a table that tabulates social class by ability. In other words, if there are twenty-five girls in the category high ability/non-manual background' then there must be twenty-five girls in every other category.

The task we are going to ask our experimental subjects to perform is to allocate the pupils to ability bands, but under three different sets of conditions.

> *Condition A*
>
> There will be as many positions in the high-ability band as there are pupils in the high-ability group, with a similar match for the middle- and low-ability bands and groups.
>
> *Condition B*
>
> There will be fewer positions in the high-ability band than there are pupils in the high-ability group and more positions in the middle-ability band. There will be as many positions in the bottom band as there are pupils in the bottom-ability group.
>
> *Condition C*
>
> There will be fewer positions in the bottom-ability band than there are pupils in the bottom-ability group, more in the middle band and equivalence in the top-ability band.

Condition A gives teachers the opportunity to use measured ability alone to distribute pupils between ability bands. Condition B forces teachers to use criteria other than measured ability to 'save' certain pupils (and not others) from the middle band. Condition C forces teachers to use criteria other than measured ability to place in a higher band pupils who might otherwise be placed in the bottom-ability band.

Our hypothesis for this experiment might be that if teachers use evidence of social class to place pupils in ability bands then:

(a) There will be no exact correspondence between measured ability and band position for Condition A, and the discrepancies will be shown in a co-variation between social class and placement in ability bands.

and/or

(b) For Condition B, teachers will show a stronger tendency to place one social class of pupil in a lower band than pupils of another social class. There will be a co-variation of social class and placement in Bands 1 and 2.

and/or

(c) For Condition C, teachers will show a greater tendency to place one social class of pupil in a higher band than pupils of another social class. There will be a co-variation of social class and placement in Bands 2 and 3.

It is worth considering the way in which this structure controls the variables. It controls for differences in ability by ensuring that for each condition teachers have exactly the same distribution of abilities within their set of pupils. In this way, different outcomes for Conditions A, B and C cannot be due to differences in the ability distribution. The structure also controls for any correlation between social class and measured ability (which occurs in real life), because each social class in the experiment contains the same ability distribution and vice versa. Differences between teachers are also controlled by making sure that each group of teachers is similar, as far as is possible. This is obviously less amenable to experimental control and is a case for running the experiment several times with different groups of teachers to check the results.

In addition, for Condition B, ability was controlled by making it impossible for teachers to use the criterion of ability alone to decide on Band 1 placements. Thus, whichever high-ability pupils they allocated to Band 2, they must have used some criterion other than ability to do so, or to have used a random allocation. Given this, we should be able to see the independent effect of this other criterion (or of other criteria) in their decisions. Much the same is true of Condition C. By having a Condition B and a Condition C we have controlled for any differences there might be in the teachers' behaviour with regard to placing pupils in higher or lower bands. By juxtaposing Condition A with Conditions B and C we can control for the effects of unforced, as against forced, choice. Our interest was in social class rather than in gender or pupils' other characteristics, but even if it turned out that teachers were basing decisions on gender or some other characteristic we have already controlled for these.

There are obviously possibilities for other confounding variables to spoil our experiment. We might, for example, inadvertently allocate more teachers vehemently opposed to streaming to one of the groups than to the others, or a disproportionate distribution of teachers with different social origins might confound our results. Such problems can never be entirely overcome.

Activity 51 (allow 10 minutes)

Suppose we ran this experiment as described and found that for all three groups there was no co-variation between social class and ability band placement: that is, no tendency for teachers to use social class as a criterion of ability band placement. What conclusions might we reasonably draw?

We might conclude that these teachers showed no social-class bias in this regard and perhaps that, insofar as they were representative of other teachers, social-class bias of this kind is uncommon. We certainly should consider two other possibilities, though. One is that the experiment was so transparent that the teachers saw through it. Knowing that social-class bias is bad practice, perhaps they did everything possible to avoid showing it. This is a common problem with experiments and one of the reasons why experimenters frequently engage in deception of their subjects. The second consideration is that while the teachers in this experiment behaved as they did, the experimental situation was so unrealistic that what happened may have no bearing on what actually happens in schools.

Both of these problems reflect threats to what is often referred to as ecological or naturalistic validity: this refers to the justification with which we can generalize the findings of the experiment to other apparently similar and, in particular, 'real-life' situations.

7.5 CORRELATIONAL RESEARCH

Our discussion of experimental method illustrates what quantitative researchers in education are often trying to do, whether by experimentation or by other means. Rather than trying to control variables by manipulating situations, many educational researchers engaged in quantitative research utilize ready-made, naturally occurring situations and attempt to control variables by collecting and manipulating data statistically. If the experimental researcher gains control at the expense of ecological validity, then correlational research gains what ecological validity it does at the expense of physical control. Naturally occurring situations are very rarely shaped so as to lend themselves easily to research. If you look back at the passage that described how we would control variables in our proposed experiment, you will see how difficult it would be to control for variables in a situation where we studied teachers who were allocating pupils to ability groups under real circumstances.

We shall illustrate the strategy used in correlational research by looking at how Stephen Ball used Beachside Comprehensive as a site for a natural experiment on ability banding. While in his Table 2.5 Ball simply displays the co-variation between social class and allocation to Band 1 or Band 2, he does not conclude from this that the allocation is biased against working-class children. He recognizes that this co-variation may be the product of co-variation between social class and ability. In order to test the hypothesis that there is social-class bias in the banding allocations, he sets out to control for ability. He does this, as we have seen, by relying on the NFER tests for reading comprehension and mathematics, which had been administered to many of the entrants to Beachside in their primary schools. He employs statistical tests to assess the relationship between social class and allocation to bands, having controlled for ability. The results of those tests are reported at the bottom of two of his tables. In this section we shall explain how he obtained his results and what they mean.

A great deal of quantitative analysis involves speculating about what the data would look like if some causal connection existed between variables (or if no causal connection existed between variables) and comparing the actual data with these predictions. This kind of speculation is, of course, a device to compensate for the fact that under naturally occurring situations variables cannot be physically controlled. The researcher is saying in effect: 'What would it look like if we had been able to control the situation in the way desired?'

We might therefore ask: 'What would data on banding and social class look like if there were no relationship at all between banding and social class: that is, if there were a *null* relationship?' Table 11 provides the answer to this question. We have collapsed the data into two categories: 'non-manual' and 'manual' (leaving those whose social class was unclassified out of account).

Table 11 Numbers of pupils of different social classes occurring in each band, observed figures (O figures), compared with numbers to be expected in each ability band if there were no relationship between banding and social class (E figures)

	Manual		Non-manual		
Band	0	E	0	E	Total
Band 1	54	70	60	44	114
Band 2	83	69	29	43	112
Band 3	20	18	10	12	30
Total	157		99		256

To create this picture we assumed that a null relationship between social class and banding would mean that pupils from different social classes would appear in each band in the same proportion as they appear in the year group as a whole. In the year group as a whole non-manual and manual pupils appear roughly in the proportions 10:16 (99:157). In Band 1 there are 114 places and sharing them out in these proportions gives us our expected figures: 44 and 70. And the other bands are similarly treated.

Comparing the O (observed) and the E (expected) figures by eye in Table 11 should show you, once again, that social class does have some role to play in the real distribution. For example, if social class were irrelevant there would be 16 fewer non-manual children in Band 1 (60 - 44) and 16 more manual children in that band (70 - 54).

As we noted earlier, this in itself does not mean that teachers are making biased decisions against working-class pupils, or in favour of middle-class pupils, in allocating pupils to bands. It remains possible that there are proportionately more middle-class pupils in Band 1 because there are proportionately more middle-class pupils of high ability. This means that the co-variation between social class and ability banding reflects a co-variation between social class and ability. We do find such co-variation in Ball's data as shown in Table 12.

Table 12 Social class and scores on NFER reading comprehension test (percentages and numbers)

Score	Working-class pupils	Middle-class pupils
115 and over	7% (4)	26% (7)
100-114	45% (26)	53% (14)
1-99	49% (29)	22% (6)
	101% (59)	101% (27)

Source: Derived from Ball (1981) p. 33 (reproduced in Offprints Reader, p. 71)

Table 12 relates only to the test of reading comprehension. You might like to check whether the same is true of the mathematics scores. Note that these scores are for a sample of 86 pupils only: those for whom the NFER data were available. We shall be commenting further on this later.

Activity 52 (allow 20 minutes)

Now try your hand at constructing a table to show what distribution of test scores would be expected if there were no relationship between social class and scores on the test for reading comprehension. Follow the procedures we adopted to construct Table 11.

When you have done this, comment on the comparison between your results and the data in Table 12, which you should have used to obtain 'observed' columns.

Your result should be similar to Table 13.

Table 13 Observed distribution (O) of test scores by social class, compared with those to be expected if there were no relationship between test score and social class (E)

	Working-class pupils		Middle-class pupils		
Score	0	E	0	E	Total
115+	4	8	7	3	11
100-114	26	27	14	13	40
1-99	29	24	6	11	35
Total	59		27		86

There is no reason why you should not have given the results in percentages, but if you wanted to make further calculations you would have to convert them back into numbers.

Comparing the observed and the expected figures by eye shows you there is co-variation between social class and test score. For example, if there were no relationship between social class and test score then there would be four fewer middle-class children and four more working-class children scoring 115 and above.

Assuming that these scores and teachers' judgements are based on ability, the distribution of children to bands should co-vary with test results. Of course, they should also co-vary with social class, since social class co-varies with test results as well.

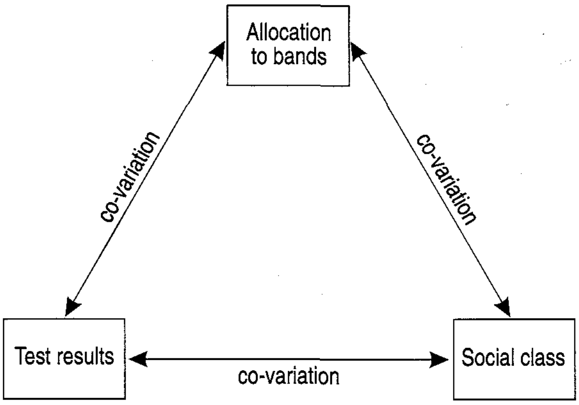


Figure 7 Co-variation among test scores, allocation to bands and social class

When everything varies together, it is difficult to judge the contribution of any particular factor. As things stand, we cannot see whether the co-variation between social class and allocation to bands is simply due to the fact that middle-class children have higher ability, as indicated by a test, or whether other social-class-related factors not associated with ability are playing **a** part. It is highly likely that both social-class-related ability and social-class-related 'non-ability' factors are at work.

As with an experimental approach, in order to tease out the relative contribution of different factors it is necessary to control variables. In correlational research, however, we do this through manipulating the data rather than the situation. What this means is that we compare cases where the variable we wish to control is at the same level or varies oniy within a small range.

Activity 53 (allow 15 minutes)

Look back at the article by Ball (Article 7 in the Offprints Reader). How does he attempt to control for the co-variation of ability with social class? What conclusion does he draw?

Ball takes pupils with the same range of test scores, but from different social classes, and investigates how they are distributed in the ability bands. This is an attempt to break out of the co-variation triangle (Figure 7 above) by holding one of its corners fixed. Thus, it can be argued that, where the test ability of the pupils is the same, any differences in band allocation that co-vary with social class must be due to the effects of social class over and above the linkage between social class and test score.

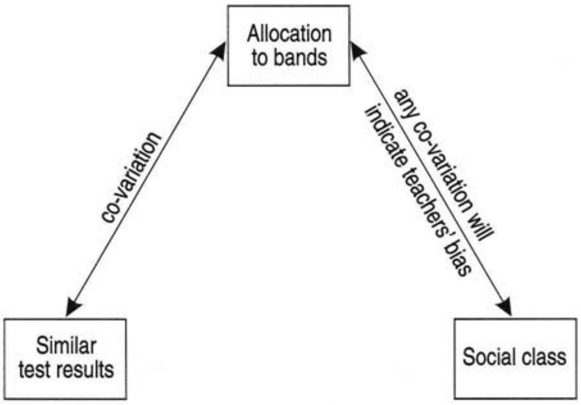


Figure 8 Controlling for test results

Ball used a statistical test to investigate the strength of the relationships amongst these factors. We shall look at the test he used later, but here we will mirror his procedures in a way that is now familiar to you.

Table 14 Observed distribution (0) of pupils scoring 100-114 in ability bands by social class, compared with that expected if social class played no part in the distribution (E)

	Working-class pupils		Middle-class pupils		
Band	0	E	0	E	Total
Band 1	10	14	12	8	22
Band 2	16	12	2	6	18
Total	26		14		40

Activity 54 (allow 5 minutes)

Looking at Table 14, how far would you say that it supports Ball's argument that even when test ability is held constant social class affects the distribution of pupils to ability bands?

So long as we take Ball's data at face value, the data in Table 14 confirm his claim. For example, if social class did not enter causally into the distribution of pupils between ability bands then there should be four (or five) fewer middle-class pupils with test scores of 100-114 in Band 1. Similarly, there should be five (or four) more working-class pupils with this kind of score in Band 1.

In our discussion of Ball's data up to now, we have often asked you to draw conclusions by visual inspection. But as we saw in Part 1, Section 3-3, statistical tests are often used to measure the differences found between observed and expected figures; in particular to assess the likelihood that they are the product of random error.

Statistical testing derives from knowledge about the laws of chance and we actually know a great deal more about chance than **we** know about non-chance occurrences. Paradoxically, perhaps, statistics applies certain knowledge about chance to the uncertainties of everything else.

We can illustrate this by simulating the allocation of pupils to ability bands at Beachside.

Activity 55 (allow 30 minutes)

Take a standard pack of cards. Shuffle and select twenty cards, ten red and ten black. Let the red cards be the children of manual workers and the black cards the children of non-manual workers. Shuffle and deal the cards into two piles. Call one pile Ability Band 1 and the other Ability Band *2.* Count the number of red cards and the number of black cards in the Band 1 pile (you can ignore the other pile because it will be the exact mirror image in terms of the number of red and black cards). You know two things intuitively: first, that the pile is more likely to contain a roughly equal number of black cards and red cards than to contain only one colour; second, that it is also rather unlikely to contain exactly five red cards and five black cards. Put another way, you know that a sample of ten cards selected by chance (a random sample) will reflect the proportions of red and black cards in the population of twenty, but that chance factors will make it unlikely that it will reflect this distribution exactly.

In fact the chances of these two unlikely events occurring can be calculated precisely assuming random allocation. Of course, shuffling cards does not give us a perfectly random distribution of the cards, but it does approximate to such a distribution. The probability of you dealing ten red cards into one pile would be one chance in 184,756. This is because there are 184,756 ways of selecting ten cards from twenty and only one of these ways would result in a selection of ten reds. On the other hand, the probability of your ending up with two piles each containing no more than six of one colour would be about 82 in 100, since there are nearly 82,000 ways of selecting no more than six of one colour.

Now imagine asking someone else to allocate the same twenty cards to the two piles, any way they like, but according to a principle unknown to you. Suppose the result is a Band 1 pile entirely composed of black cards. This raises the suspicion that in allocating cards to piles they showed a bias for putting black cards in the Band 1 pile and red cards in the Band 2 pile. Knowing what you know about the distributions likely to occur by chance, you can put a figure to the strength of your suspicion about their bias. You could argue as follows. If a set of ten red and ten black cards are shuffled and divided into two piles at random, then the chances of an all-black (or an all-red) pile are around 0.001% (actually 1 in 92,378). Therefore it is very unlikely that this pile resulted from an unbiased (random) distribution.

If the result was a pile of six black and four red cards then you could have argued that distributions with no more than six of one colour could have occurred by chance 82% of the time. The actual distribution might have been due to a small bias in favour of putting red cards in the Band 1 pile, but the most sensible conclusion for you to reach would be that there is insufficient evidence for you to decide whether this was a biased distribution or an unbiased, random distribution.

The situation faced by Ball was very similar to this example of card sorting. He was suspicious that pupils (our cards) were being sorted into ability bands (our piles) in a way that showed bias against working-class pupils (red cards) and in favour of middle-class pupils (black cards). To check out this suspicion he used a statistical test that compares the actual distribution (the distribution in our experiment obtained by unknown principles) with a distribution that might have occurred by chance. The result of using the test is a figure which will show how reasonably he can hold to his initial suspicion.

You have already encountered the way in which the figures were set u**p** for statistical test by Ball in our Table 14. We suggest that you now look **a**t this table again. Remember that for this test Ball has selected pupils from within the same range of measured ability, so that he can argue that any differences in allocation to bands are likely to be due to social-class bias, or chance: in other words, he has controlled for ability. The statistical test will help to control for the effects of chance. Therefore, logically, to the extent to which the actual figures depart from what might have occurred by chance, this is likely to be due to social-class bias. You can now regard the 'expected' figures as the figures most likely to have occurred by chance. They are the equivalent of our 5:5 ratio of black to red cards in the example above and, in this case, are just the distribution that would be expected if working-class children and middle-class children had been allocated to bands in the same proportions as they appeared in a total 'pack' of forty.

7.6 THE CHI-SQUARED TEST

The statistical test Ball used is the chi-squared test, pronounced 'ki-squared test'. (You will often see this written simply as 'the χ^2 test', using the Greek letter chi). What the test does mathematically is to compare the actual or observed figures (O) with the figures expected by chance (E). The comparison is done by subtraction.

There are four sets of O and **E** figures in Table 1**4** and so there are four sets of subtractions. These are squared and divided by the relevant **E** figure. The results are then added together. This figure is looked up in a ready reckoner called a 'Table of critical values for χ^{2}'. It tells us how often a particular deviation from what might have occurred by chance, would have occurred by chance. Note the double use of chance in this sentence. Like our 5:5 ratio in the card-sorting simulation, the E figures are the single most likely chance figures, but other chance combinations may occur. We want to know how often.

One feature of the chi-squared calculations makes them look more complicated than they are and tends to obscure what is going on. This is that the differences between O and E figures are squared.

Doing this avoids the problem that arises from the fact that if you add up all the differences they would total to zero, cancelling each other out. Squaring them converts negatives to positives and leaves you with all positive numbers which express the amount of variety (variance) in the data.

Statistical calculations are full of squarings and the taking of square roots mainly for this reason. It is probably one reason why non-mathematicians find statistics so threatening. In fact, though, what is being done is quite simple.

Table 15

	Working-class pupils		Middle-class pupils		
Band	0	E	0	E	Total
Band 1	10	14.3	12	7.7	22
Band 2	16	11.7	2	6.3	18
Total	26		14		

The calculation of chi squared for Table 15, in the way Ball calculated it, is as follows:

$$\chi^2 = \frac{(10 - 14.3)^2}{14.3} + \frac{(12 - 7.7)^2}{7.7} + \frac{(16 - 11.7)^2}{11.7} + \frac{(2 - 6.3)^2}{6.3}$$

$$= \frac{18.49}{14.3} + \frac{18.49}{7.7} + \frac{18.49}{11.7} + \frac{18.49}{6.3}$$

$$= 1.29 + 2.40 + 1.58 + 2.93$$

$$\chi^2 = 8.2$$

Before you can look this figure up in a table of critical values it is necessary to work out what are called 'degrees of freedom'. These express the amount of free play in the data. In our card sorting exercise cards could only be black or red and could only be sorted into one pile or the other. If you imagine sorting the cards such that ten cards were allocated to the first pile, then by the time this was complete everything about the second pile was decided. There is onfy one degree of freedom here. If there are more degrees of freedom then there is more free play for chance and this has to be taken into consideration. The usual way of calculating degrees of freedom for a chi squared calculation is by the formula:

$$\text{df} = (\text{number of columns} - 1) \times (\text{number of rows} - 1)$$

In the simulation and in Table 15, there are two rows and two columns, thus

$$\text{df} = (C - 1) \times (R - 1) = 1 \times 1 = 1$$

Table 16 Part of the table of critical values for $\chi 2$

	Levels of significance (probability, p)					
Degrees of freedom	0.2	0.10	0.05	0.02	0.01	0.001
df= 1	1.64	2.71	3.84	5.41	6.64	10.83
df = 2	3.22	4.60	5.99	7.82	9.21	13.82

Ball's analysis had one degree of freedom and the value of chi squared was 8.2. This value exceeds 6.64, but falls below 10.83. Therefore, from Table 16, the level of probability, *p*, is <0.01.

It is perhaps more meaningful to multiply 0.01 by 100 and then to say: less than once in 100 times would this result occur by chance. Remember how the question was posed in setting up the chi-squared test. Ball asked how often would a distribution like the one in his table occur by chance. Once in a 100 is rather unlikely[17] Thus if all the other procedures followed by Ball, and repeated by us, are correct, we can be fairly confident that the observed figures in the table were not the product of chance.

A small annoyance with regard to the chi-squared test is something called 'Yates' correction' which entails reducing by 0.5 each observed figure that is greater than expected and increasing by 0.5 each observed figure that is less than expected. This should be used whenever there is only one degree of freedom. Ball should have done this in his calculation, but did not, and we followed his procedures here. If we apply Yates' correction, the first two calculations for Table 15 should have been

$$\frac{(10 + 0.5 - 14.3)^2}{14.3} \text{ and } \frac{(12 - 0.5 - 7.7)^2}{7.7}$$

and the resulting figure for chi squared should have been $\chi^2 = 6.4$.

Activity 56 (allow 10 minutes)

Try your hand at using the table of critical values (Table 16) to establish the level of significance for a value of chi squared of 6.4. There is still of course only one degree of freedom.

On this basis, how many times out of 100 would a distribution like the one in Ball's table occur by chance?

You should conclude that it is still rather unlikely. The figure is 0.02, or twice in a hundred.

[17] Conventionally, statisticians do not take seriously any level of probability greater than 0.05; in other words, where you would expect the observed pattern to occur by chance 5 times or more in every hundred.

Activity 57 (allow 20 minutes)

If you are unsure about the workings of chi squared, do your own calculations for the card sorting simulation where the observed figures were Band 1,10 black cards, and Band 2, 10 red cards. Remember to take away Yates' correction, since there is only one degree of freedom here.

Our answer to Activity 57 is calculated from Table 17.

Table 17

	Black		Red	
Band	0	E	0	E
Band 1	10	5	0	5
Band 2	0	5	10	5

$$\chi^2 = \frac{(10 - 0.5 - 5)^2}{5} + \frac{(0 + 0.5 - 5)^2}{5} + \frac{(0 + 0.5 - 5)^2}{5} + \frac{(10 - 0.5 - 5)^2}{5}$$

$$\chi^2 = 16.2$$

Reading from Table 16, since the degree of freedom is one, $p < 0.001$. In other words, less than once in 1000 times would this result occur by chance and you can be fairly certain that the way the cards were actually distributed was not randomly.

7.7 RELATIVE CONTRIBUTIONS: MEASURES OF THE STRENGTH OF ASSOCIATION

Both ability, as measured by a test score, and social class, as measured by whether the occupation of the head of a household was manual or non-manual, seem to have some independent effect on the way in which pupils are distributed to bands at Beachside as represented by Ball's data. As yet, we have not established which has the strongest influence. To estimate this there are various statistical techniques available that go under the general heading of 'correlation co-efficients'.

We shall demonstrate the use of **a** measure called phi, represented by ø, which is easily calculated from the value of chi squared. The minimum value of phi is 0 and the maximum value (for 2x2 tables) is 1. Ball used the co-efficient *C*, which is also easy to calculate, but is difficult to interpret because its maximum value is less than 1 and varies from one data set to another.

Unlike Ball, who used oniy a part of the data available to him, we shall use all the data available. You will see that below we have made separate calculations of chi squared for:

social class and banding, irrespective of ability (Table 18);

ability and banding, irrespective of social class (Table 19);

social class and ability (Table 20).

From this we shall be able to see which are the strongest relationships in the causal network.

Table 18 Social class and banding

	Working-class pupils		Middle-class pupils		
Band	0	E	0	E	Total
Band 1	18	26	20	12	38
Band 2	41	33	7	15	48
Total	59		27		86

chi squared = 12.3, df = 1,p < 0.001

Table 19 Test score and banding

	100+		1-99		
Band	0	E	0	E	Total
Band 1	32	22.5	6	15.5	38
Band 2	19	28.5	29	19.5	48
Total	51		35		86

chi squared = 15.82, df = 1,p< 0.001

Table 20 Test score and social class

	Working-class pupils		Middle-class pupils		
Test score	0	E	0	E	Total
100+	30	35	21	16	51
1-99	29	24	6	11	35
Total	59		27		86

chi squared = 4.53, df =1, $p < 0.02$

Tables 18 and 19 show you what you already suspected from working with the data earlier. They show that there is a statistically significant relationship between social class and banding and between ability (test score) and banding. The chi-squared test tells you that these patterns are most unlikely to have occurred by chance. In both cases the probability of this happening is under 1 in 1000 ($p < 0.001$). You will note that chi squared for Table 19 is higher than that in Table 18. This is an indication that allocation to bands shows a stronger relationship with test scores than with social class.

Comparing the figures for chi squared can be misleading, however. A better measure of the relative strength of the relationship is derived from calculating phi, ϕ The formula for this is

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

This means that we must divide chi squared by the number in the sample and take the square root of the result.

To read phi or any other correlation co-efficient as a measure of the strength of a relationship, it is conventional to square it and multiply by 100 to produce a percentage figure. Thus:

Phi for social class and banding (Table 18) is 0.378
or 14% (0.378^2 x 100 = 14.288).

Phi for test score and banding (Table 19) is 0.429 or 18%.

Phi for social class and test score (Table 20) is 0.230 or 5%.

One way of looking at this is in terms of Figure 9.

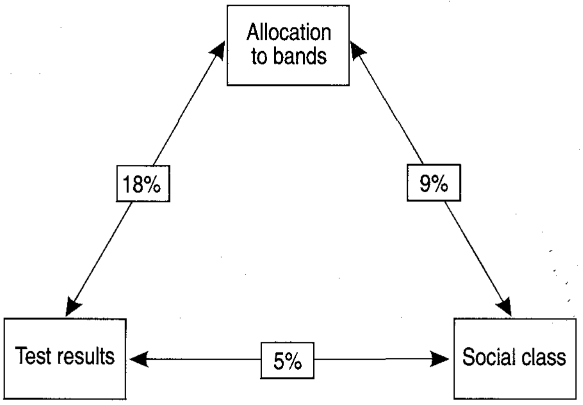


Figure 9 Strength of association between tested ability, social class and banding in a causal network

In Figure 9 we have subtracted the 5% for the correlation between ability and social class from the 14% for the correlation between social class and allocation to bands. This is a way of separating the influence of social class itself from the influence of social-class-related ability. We do not have to add that 5% to the influence of ability alone because it is already included within the 18%.

It is important to be clear about what the percentages in Figure 9 mean. It might be tempting to look at Figure 9 and say that 18% of the pattern in banding is caused by ability, as measured by test scores. Statisticians often come close to this kind of statement by using phrases such as 'explained by' or 'accounted for by'. This is slightly misleading. You already know that test scores do not cause banding in any simple way, although you might suspect that some underlying ability of pupils causes their test score and causes teachers to do what they do to allocate pupils to ability bands.

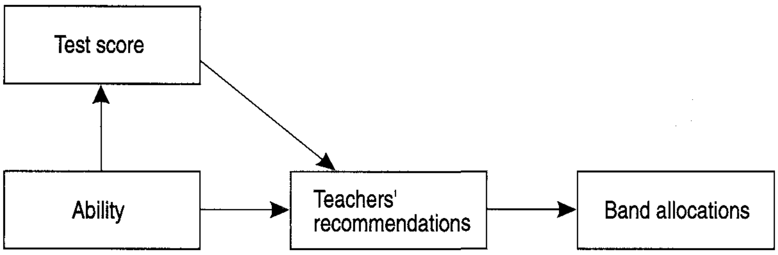


Figure 10 The possible role of ability in the causal network

Correlations may or may not indicate causes, but what they always do represent are predictions. Thus when we say that '18% of the distribution of pupils in the banding system can be accounted for by the distribution of test scores', what we actually mean is that knowing the distribution of test scores improves our ability to predict the distribution within ability bands, and it does this by 18%. This can be demonstrated quite easily.

Suppose all you knew about the distribution of pupils within bands was that there were 38 in Band 1 and 48 in Band *2* (Table 21). The best prediction you could make for any particular pupils would be that they would be in Band 2. This is simply because there are more pupils in Band 2 than in Band 1. If you guessed that all the pupils were in Band 2 you would in fact be right 48 times out of 86 — a success rate of 56%.

Table 21 Test score and banding

Band	100+	1-99	Total
Band 1	32	6	38
Band 2	19	29	48
Total	51	35	86

If you were told the distribution of pupils across bands according to their test score, you could improve your prediction. Now your best bet would be that all pupils with test scores above 100 would be in Band 1 and all pupils with test scores from 1 to 99 would be in Band 2. You would be correct on 6l occasions (32 + 29) out of 86 and your success rate would have increased to 71%.

Thus, knowing the test scores improves your ability to predict by 15%: that is, it reduces the error in guessing by 15%. (In the first instance the percentage error is 100-56=44%, in the second it is 100-71=29% and 44-29=15%.)

What phi and other correlation co-efficients do is to provide a more exact measurement of the extent to which the distribution of one variable (say, allocation to bands) can be predicted from knowledge of the distribution of another (say, test score). The 18% in Figure 9 says that knowing the distribution of test scores we can improve our ability to guess the distribution into bands by about 18%. (This is not quite the same estimate as the 15% we gained by guessing. Different methods produce different estimates.) The fact that we can improve our predictions in this way is *prima-facie* evidence that the two variables are linked together in some causal network. Exactly what is the nature of that relationship, however, is something which has to be inferred.

7.8 REPRESENTATIVE SAMPLES AND STATISTICAL TESTING

As we noted earlier, Ball used data on 86 pupils to examine or illustrate what was happening to all pupils. So the question arises of how representative the allocation of the sample to bands is of the allocation of the *whole year group to* bands. The chi-squared test does not inform us directly about this. It tells us is that if the teachers had allocated an infinite number of working-class and middle-class children to the bands on the basis of their measured ability, the likelihood of drawing a sample of 86 children from this population which showed the inequalities in distribution Ball found is very small. As it stands, then, this is evidence to suggest that the teachers did not allocate these children to bands on the basis of measured ability alone. However, in using any test we need to know the conditions the data must meet for it to work properly. In the case of the chi-squared test there are two of these: the samples drawn of working-class and middle-class children must be both independent and random. Ball's use of chi squared meets the first of these conditions (the composition of each social-class sample did not have any effect on that of the other), but it clearly does not meet the second. Ball's sample was determined by the test result data he could obtain. While it was not selected systematically, neither was it selected randomly. Therefore, there is a risk that the difference between observed and expected allocations *is a.* product of factors involved in the way in which Ball's sample was drawn from the year group.

This leads us on to a further point. Since we are interested in whether allocation was biased in the year group as a whole, rather than just in the sample, it is important to note that the latter *is* not very representative of the year group, as regards the allocation of working-class and middle-class students to bands (taking no account of ability). If we compare the data for the sample with that in Table N2 of Ball's article, we find that by comparison with the whole year group both middle-class pupils in Band 1 and working-class pupils in Band 2 are over-represented in the sample. As a result, there is a smaller percentage of middle-class pupils in Band 2 and of working-class pupils in Band 1 in the sample than in the school year. Ball's argument is that teachers are biased in favour of middle-class pupils and place more of them in Band 1, and therefore place more working-class pupils in Band 2. His sample is, however, already biased in this direction. The extent of the bias is more easily seen if we calculate the figures for a precisely representative sample and compare them with the actual sample, as in Table 22.

Table 22
Comparing Ball's sample with the numbers expected in a representative sample

	Working-class pupils		Middle-class pupils		
Band	actual	representative	actual	representative	Total
Band 1	18	21	20	23	38
Band 2	41	32	7	11	48
Total	59	53	27	34	86

chi squared = 3.92, df = 1, $p < 0.05$

From the chi-squared test you can see that a sample that deviated from the representative sample to the same degree as Ball's might be drawn at random less than 5 times out of a 100. Put the other way, around 95 samples out of 100 drawn at random would be more representative than Ball's sample. Clearly, Ball's is not a very representative sample in these terms. And we also know that the direction in which it deviates from being representative is precisely the direction which Ball takes as evidence of bias in the allocation process. Of course this could be because the distribution of higher and lower ability pupils across the social classes in Ball's sample is very different from that in the year group; but we have no way of gauging this. This does not disprove the validity of Ball's chi-squared result, but it should make us rather cautious about accepting it as evidence that there was bias in the teacher's allocation of pupils to bands in the year group as a whole.

7.9 STATISTICAL SIGNIFICANCE AND SUBSTANTIVE SIGNIFICANCE

This is not the end of the story. It is quite possible, and indeed quite common, to work out statistical tests correctly and to obtain statistically significant results, but to demonstrate nothing of any importance in a theoretical or practical sense. Worse still, it is possible to produce results that are misleading!

There is an important substantive matter which Ball neglects, which could have threatened the validity of his findings. To recapitulate, Ball's main hypothesis was that social-class bias was at play in allocating pupils to ability bands. To demonstrate this bias he performed a chi-squared test, which compared the actual distribution of pupils of different social classes into bands with the distribution that would have occurred if the two classes had been allocated to bands proportionately (that is, if pupils of the same ability had been allocated to bands randomiy without taking their social class into consideration). Deviation from a proportional distribution is how Ball has operationalized 'social-class bias'. The question we should ask is whether this is a sensible way of operationalizing social-class bias.

For his demonstration Ball only used the figures for pupils with test scores of 100—114. What he did not take into consideration was the fact that, in reality, the decisions on band allocation for these particular pupils are influenced by the allocation of pupils with test scores of over 114. This is because allocations are usually made in a situation where the number of places in bands is more or less fixed.

Let us use playing cards again to simulate the situation. We shall still have ten red cards and ten black cards and once again ten cards will have to be placed in each pile. This time, however, there will be four black kings and one red king. We shall introduce the rule that all kings must be allocated to the Band 1 pile first and then the other cards divided into the two piles at random. You will see immediately that this considerably reduces the chances of red cards appearing in the pile representing Band 1. Whether it does so 'unfairly' depends on our judgement about the rights of kings to a place in that pile.

At Beachside 'kings' are the pupils scoring 115 and above. Within the framework within which Ball is operating, achievement-test scores are taken to represent the ability to benefit from placement in different bands. There really seems to be no objection to assuming that those pupils with high ability in these terms should have priority claims to places in Band 1. In Ball's sample, there are more middle-class pupils with high ability than working-class pupils. By giving high-ability pupils priority for Band 1 placement, we automatically give priority to middle-class pupils.

Taking this into consideration, we can work out a 'fair' distribution for ability banding assuming that the number of places in each band is fixed. In terms of the reading comprehension test (see Table 2.6 in Ball's article), there are eleven pupils of the highest ability (that is, those with scores of 115+), four working-class and seven middle-class pupils. Give them Band 1 positions. Since there are thirty-eight Band 1 positions there are now only twenty-seven places left. There are forty pupils with test scores of 100—114 with an equal claim to Band 1 positions. To avoid social-class bias we shall distribute them in proportion to the number of working-class and middle-class pupils. Since there are twenty-six working-class and fourteen middle-class pupils, we shall give Band 1 positions to 17.5 working-class pupils and 9.5 middle-class pupils (you can do things with statistics which you could never do in reality!). Now there are 4 + 17.5 working-class pupils and 7 + 9.5 middle-class pupils in Band 1. The remainder of the pupils are now allocated to Band 2. These procedures provide us with the expected figures given in Table 23.

Table 23 Distribution of pupils of different social classes in ability bands. Observed distribution compared with distribution expected from 'fair' allocation taking highest ability pupils into consideration

	Working-class pupils		Middle-class pupils		
Band	O	E	O	E	Total
Band 1	18	21.5	20	16.5	38
Band 2	41	37.5	7	10.5	48
Total	59		27		86

chi squared = 2.06, df = 1, $p < 0.10$

From the figures in Table 18, the extent to which the actual figures departed from a simple proportional distribution was calculated. Chi squared was then 12.3, with a very high level of statistical significance. Our new calculation compares the actual distribution with a distribution that might have occurred had pupils been allocated to bands by giving those of the highest ability priority to Band 1 and then distributing the remainder proportionately between pupils of different social classes. We think this is a better model of a 'fair' distribution than the one adopted by Ball. Now chi squared is only 2.06. It is statistically significant only at the 10% level. This means that had teachers at Beachside really been distributing pupils to bands on the same basis as our model, then 10% of random samples of 86 drawn from the whole year population would show this degree of 'social-class bias', simply because of the chancy nature of samples. As we noted earlier, the convention in statistics is not to take as significant any level of probability greater than 0.05, or 5%.

Put more formally our null hypothesis was that there was no statistically significant relationship between social class and allocation to bands when ability was controlled. Following statistical convention, we would want $p < 0.05$ to reject the hypothesis. In our calculations p did not reach the 5% level, hence our null hypothesis is not rejected and we have no good grounds for saying that social-class bias influences decisions on band allocation. For this reason, over and above the fact that he may have been working with an unrepresentative sample, Ball does not provide convincing evidence that social-class bias is an important influence on the distribution of pupils to bands in the cohort of pupils he studied at Beachside. It is worth emphasizing that this does not necessarily mean that there was no social-class bias, merely that Ball gives no strong evidence for it. In the conclusion to his book, Ball himself describes the evidence as 'generally inconclusive' (Ball, 1991, p. 283), though on somewhat different grounds.

7.10 UNDERLYING ASSUMPTIONS

It is important to be as critical of our own procedures as we have been of Ball's. The question must be asked: how reasonable is our model of a fair distribution? We think that it is an improvement on Ball's model, for two reasons. First, it takes into consideration the claims of those with high test scores to be placed in the top band and, secondly, the fact that the number of places in the top band is likely to be restricted. Our application of the model, however, does give rise to some problems. We adopted the model on the assumption that the placement of any one pupil is likely to result from all the decisions about all the pupils. Unfortunately, we only know about the test scores of the pupils in Ball's sample and, as we know, that sample is unrepresentative. Our application of the model to Ball's data relies on the assumption that in the cohort as a whole the ratio of high to middle to low test scores is 11:40:35. In other words, we assume it is similar to that in the sample. We also rely on the assumption that test scores and social class in the cohort covary in the same way as in the sample. If, in fact, there were just as many working-class pupils as middle-class pupils with test scores of 115 and above in the year group as a whole then our procedures would be severely awry All we can say about this is that when NFER tests are administered to large groups of pupils, middle-class pupils do usually score more highly on average.

It is also worth remembering at this point an issue which we raised earlier about Ball's operationalization of ability. We can ask: 'Why should we regard NFER achievement test scores as a better measure of a pupil's ability than the estimates of junior-school teachers?' After all, the bands group pupils for all subjects, whereas the tests measure achievement in specific areas. While there are very few ways of getting a high score on a test, there is a large number of ways of getting a low score, many of which would not be indicative of underlying inability. We should rightly complain if junior-school teachers did not take this possibility into consideration in making recommendations about the allocation of pupils to bands in secondary school. For what it is worth, Ball's data do show more pupils being 'over-allocated' than 'under-allocated'. Ball's argument was constructed on the basis that test scores represented 'true ability', so decisions on band allocation that departed from what might be predicted by a test score could be seen as social-class bias. Our argument has been that Ball has demonstrated no social-class bias, but in order to pursue it we have had to assume with him that test scores provide a sufficient indication of the way in which pupils should be banded by ability. This is not an assumption we should like to defend.

Activity 58 (allow 6 hours)

1. At this point we would like you to produce a summary of the various questions that we have raised about Ball's analysis of his data. This will provide a basis for carrying out your own assessment of other similar articles.
2. When you have done this, you should read 'Figuring out ethnic equity: a response to Troyna' by Roger Gomm (Article 12 in Reader 2). This is a critical assessment of the article by Barry Troyna that we discussed in Part 1, Section 3.3. We are not able to provide you with Troyna's article for copyright reasons, but Gomm gives a clear outline of the claims Troyna makes. You should find it easy to follow Gomm's assessment on the basis of the work you have done so far. Make a note of the major points Gomm makes. How far do you agree with his assessment?

7.11 REGRESSION ANALYSIS

Up to now we have concentrated on the underlying logic of quantitative data analysis, introducing a small number of techniques as and when they became appropriate for the analysis of the data from Ball's study. You will find in the literature quite a lot of research that relies primarily on the techniques we have discussed so far. However, they represent oniy a very small range of the statistical techniques that are available and that have been used by educational researchers. We cannot hope to cover all the others, but what we shall do in the remainder of this section is to introduce one of the most frequently used of the more advanced techniques, 'regression analysis'. We shall illustrate its use in multi-level modelling, an approach which is designed to discover the contribution that schools make to the achievement levels of their pupils. As you will see, regression analysis is a development of the techniques to which you have already been introduced. Before we discuss it fully, however, we need to cover one or two other issues.

Types of data

The kinds of statistical tests that can be employed in educational research depend on the kinds of data that are used. So far we have been using data as if they were nominal or categorical in character. This is the form of data that is least amenable to statistical use. For regression analysis, higher levels of data are required. Below we outline the standard classification of different types of quantitative data.

- *Nominal-level data*

  Examples: classifications of gender and ethnicity. This type of data is organized in terms of categories which are mutually exclusive and exhaustive. Thus, in the case of gender and ethnicity it should be possible to assign all people to one and oniy one category. It is important, though, to recognise the limits of this kind of data: you cannot add together the number of males and the number of females, divide by the total and come up with an 'average gender'. You cannot multiply ethnic groups together and come up with something different. Within the bounds of common sense, however, you can collapse nominal categories together. For example, Ball collapses the Registrar General's classes into the two-category system, of manual and non-manual.

- *Ordinal-level data*

  Examples: ability bands or the rank order of pupils in a form. These are ordinal-level data in the sense that they can be ranked from highest to lowest. You cannot, however, give a measurement for how much higher Band 1 is than Band 2. In other words, ordinal data can be ranked, but the intervals between the ranks cannot be specified.

- *Interval-level data*

  Interval data have a standard and known interval between points on the scale. Thus, in the case of SATs scores, if we were justified in assuming that the difference between a Grade 2 and a Grade 6 is the same as that between a Grade 6 and a Grade 10, then we could say with justification that SATs scores are interval data.

- *Ratio-level data*

  Chronological age, parental income, height, distance travelled to school, would all be ratio-level data. This is because there is a standard scale of measurement that can be used which has both equal intervals and a true zero point. With ratio-level data we know that a score is a specified number of equidistant units away from zero. Although this is not precisely true of GCSE grades or IQ scores, many educational researchers behave as if it is.

Activity 59 (allow 10 minutes)

What level of data are the following:

(a) pupils classified into the social-class groupings in Ball's Table 2N;

(b) NFER test scores grouped into three categories;

(c) NFER test scores showing the actual score per pupil;

(d) examination marks;

(e) position of a pupil in the banding system;

(f) social class in two categories: working class and middle class..

Here are our answers:

(a) Ordinal-level data, if you ignore the problem of the unclassified pupils, but more safely regarded as nominal.

(b) Technically this is interval- or ratio-level data, but clumped together in this way it is not much of an improvement on ordinal-level data.

(c) There is some debate about whether test scores are interval- or ratio-level data. In most educational research they would be treated as ratio-level data.

(d) The answer will depend on the way in which the examinations are marked, but it would be safe to assume no higher than interval-level data.

(e) Ordinal-level data.

(f) Nominal-level data.

These answers indicate what is the highest level at which you can use each kind of data, since you can always use data of a higher level as if it were of a lower level And, given uncertainty about whether particular data meet the requirements of higher levels of measurement, sometimes this is advisable.

Diversity and size of sample

If you remember that in quantitative work we are usually looking for co-variation, you will understand that co-variation is most easily seen where data can be scaled precisely, as with interval-level and ratio-level data. We can illustrate this speculatively by thinking about Ball's data on the allocation to ability bands of the group of pupils with test scores of 100-114 (Table 14). As his data stand, pupils with a wide range of test scores are grouped together in **a** single category. This lumping together seems to make it reasonable to assume that all these pupils should be treated similarly, but what if we knew their individual scores; what if we had interval- or ratio-level data for these pupils? Table 24 shows two possible distributions of the pupils' scores.

Table 24 Two possible distributions of NFER test scores among those scoring 100-114: by social class

	Distribution 1			Distribution 2	
Scores	Working-class pupils	Middle-class pupils	Scores	Working-class pupils	Middle-class pupils
114		2	114	3	2
113		1	113	4	1
112		3	112	2	0
111		0	111	1	0
110		3	110	2	0
109		1	109	1	0
108		1	108	2	0
107	3	1	107	2	0
106	2	0	106	1	1
105	3	0	105	1	2
104	2	0	104	1	2
103	3	0	103	2	0
102	2	0	102	1	1
101	3	1	101	1	2
100	1	1	100	0	3

Activity 60 (allow 15 minutes)

What important difference is there between these two distributions from the point of view of Ball's analysis?

On the first hypothetical distribution, placing all pupils with scores of 100-114 in Band 1 would give the actual social-class differences shown in Ball's table (Table 14). This is quite 'fair', too, because middle-class pupils are, on average, scoring higher marks than working-class pupils. In the case of the second hypothetical distribution, however, what is shown in Ball's data cannot be consistent with an even-handed treatment of pupils from different social classes because, on average, working-class pupils are scoring higher. What the actual situation was, of course, we cannot know because of the way in which differences within the groups have been submerged by treating all those scoring 100—114 as the same.

Given that the more data can be differentiated the better chance there is of accurately displaying co-variation, why do authors often work with data that are grouped into very crude categories? There are two main reasons for this. First, highly differentiated data are often not available. This is frequently the case when a researcher relies on second-hand data. For example, we might guess that the oniy test data Ball could obtain was clumped into thiee test-score categories. The second common reason relates to the size of the data set. Again, a pack of cards provides a convenient demonstration.

As you saw, a random sample of ten cards drawn from a pack of ten reds and ten black cards gave you a fairish chance of getting a sample which represented the colour distribution of the cards in the pack. That is, on 82 occasions out of 100 you might have dealt ten cards with no more than six cards of any colour. Suppose now that the 'population' of twenty cards had five hearts, five diamonds, five clubs and five spades. You should realize intuitively that the chance of a deal of ten accurately representing suits is much less than the chance of it accurately representing colours. In turn, the chance of a deal of ten accurately representing the denominations of the cards is even less than the chance of accurately representing the suits.

An adequate size for a sample is determined by the amount of the diversity you want to represent. By the same token, the smaller the sample the less diversity you can represent. Let us think of this in terms of two of Ball's important variables: social class and test scores. Ball had data that would enable him to subdivide pupils into social classes in the six categories of the Registrar General's scheme plus an 'unclassified category' (see Ball's Table N2).

Let us imagine that he had test-score data enabling him to divide pupils into groups that each represented an interval often test-score points (for example, 1-10, 11-20 ... 121-130). This results in thirteen categories, producing a table with seven columns (for social class) and thirteen rows (for scores). This gives a table with ninety-one cells: in fact, a table of around the same size as Table N2. Given that Ball onfy had test-score data for eighty-six pupils, had he subdivided as above, many of the cells in his table would have remained empty. It is probable that many of the others would have contained oniy one or two entries. In addition, Ball was interested in how pupils with particular scores from particular social classes were allocated to the three bands. This in effect makes a table with 91 x 3 - 273 cells. Even if Ball had had test-score data for the entire 296 pupils in the year group (a 100% sample) it would have been too small to display adequately the relationships between social class, test score and banding at this level of detail.

Under these circumstances it may be necessary to collapse data into cruder categories. The payoff is that patterns can be seen in the collapsed data which are not visible in highly differentiated data. On the other hand, the costs of degrading data are that important differences may become invisible and that patterns may emerge that are largely the result of the way the data have been collapsed.

The technique of regression analysis

Much sophisticated quantitative work in educational research uses the technique of regression analysis. This requires data to be of at least interval level. To illustrate what is involved in regression analysis we have invented a set of data that might apply to a class in a top-ability band in a school like Beachside. Imagine that we have data from an NFER test in English conducted in the last year of junior school, and from an examination that pupils sit at the end of their third year in secondary school. Social class is collapsed into the two categories, middle and working class, as with Ball's data.

We are going to use the data in Table 25 to investigate whether social-class differences in achievement in English increase, decrease, or stay the same over the first three years of secondary schooling. Ability in English at the beginning of the period is measured by the NFER test and, at the end, by the score in the school's English examination. For the time being ignore the column headed 'residual'.

Table 25 Fictional data set showing NFER test scores and examination results for working-class (W) and middle-class (M) pupils

Pupil	Social class	NFER test score	Examination mark	Residual
1	M	127	80	2.97
2	M	128	79	0.66
3	M	116	76	13.38
4	M	114	75	15.01
5	M	125	74	-0.41
6	W	130	68	-12.96
7	M	111	67	10.94
8	M	114	65	5.01
9	W	100	64	22.35
10	M	110	60	5.25
11	W	106	58	8.49
12	M	105	57	8.80
13	M	110	56	1.25
14	M	113	55	-3.68
15	M	104	54	7.11
16	W	101	53	10.04
17	W	105	50	1.80
18	M	100	48	6.35
19	M	100	46	4.35
20	W	103	44	-1.58
21	M	102	42	-2.27
22	W	112	41	-16.37
23	W	107	38	-12.82
24	W	100	36	-5.65
25	W	94	34	0.22
26	W	100	33	-8.65
27	W	110	32	-22.75
28	M	92	30	-1.16
29	W	87	28	3.39
30	M	100	24	-17.65
31	M	96	15	-21.40

Before we do anything more sophisticated with these figures it is worth manipulating them to see what patterns can be made visible. Since we can treat the test scores and the examination results as interval-level data, we can calculate averages (in the form of means) for the two social classes of pupils.

Averages

There are three kinds of average: the 'mode', the 'mean' and the 'median'. They all express a central point around which the scores cluster and, hence, are called measures of the central tendency.

The mode is the most frequently occurring score for a group. In our data there are no modes for examination results, since each pupil has a different score, but the mode for the test scores for all pupils is 100, which appears six times. The mode is not a particularly useful measure statistically; it corresponds roughly with what we mean verbally by 'typical'. In the case of our data it is not very useful as an indication of the central tendency, since twenty-one out of thirty-one pupils scored more than the mode. You should realize that if we treat data at the nominal level, the mode is the only kind of 'average' available.

The mean, or mathematical average, is a more familiar measure. It is calculated by adding together all the scores and dividing by the number of scores.

Table 26 Means for NFER test scores and examination results: working-class and middle-class pupils: derived from Table 25

	Test scores	Examination results
middle-class pupils	109.28	55.72
all pupils	107.16	51.03
working-class pupils	104.23	44.54

We shall not be using the third type of average, the median, in the calculations to follow, so we shall leave discussion of it until later.

Measures of dispersion

The mean by itself can be misleading, because it does not take into consideration the way in which the scores are distributed over their range. Thus, two sets of very different scores can have very similar means. For example, the mean of (2,2,4,4,6,6) and (2,2,2,2,2,14) is 4 in both cases.

In effect, you have already encountered the problem of using mean scores for statistical calculations. The problem we noted as arising from lumping together all pupils scoring 100—114 is of much the same kind.

In our data, the range of examination scores for working-class children is 40 and for middle-class pupils it is 65. (This is calculated by finding the difference between the highest and lowest scores.) Moreover, the way in which the scores are distributed across the range is very different. This can be made visible by collapsing the examination scores from Table 25 into intervals of ten (Figure 11).

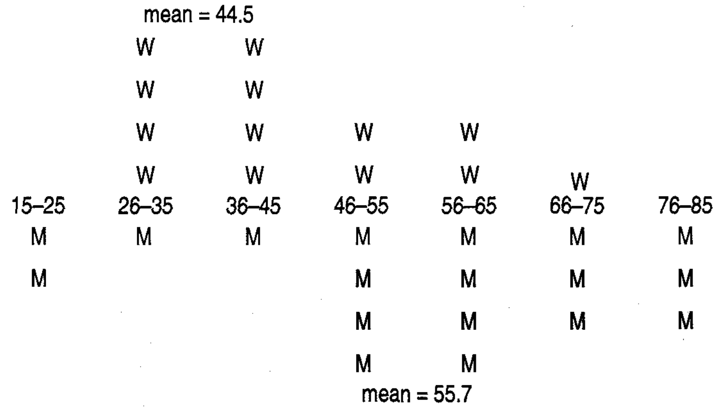


Figure 11 A comparison of the two distributions for working-class (W) and middle-class (M) pupils

You can see from Figure 11 that the scores for working-class children are clustered much more towards their mean, while the scores for middle-class children are much more dispersed. It shows that the mean for working-class pupils gives us a rather better sense of the central tendency of their scores than the mean for middle-class children does for their scores. Simply comparing mean scores for both groups would ignore this.

Standard deviation You probably realize that the range of scores also does not capture the dispersion of data very effectively. It tells us the difference between the end points of the distribution, but nothing about how the data are distributed between those points. More effective is the 'standard deviation'. To understand the idea of standard deviation you may find the following analogy useful.

In the game of bowls the best player is decided by whosoever manages to place a single wood closest to the jack. Imagine that we want a more stringent test of which of two players is the better, taking into consideration the position of all woods in relation to the jack. Thus, for the first bowler we measure the distance between the jack and each of that player's woods and divide by the number of woods. We now have a single measure of the extent to which all the first player's woods deviated from the position of the jack: an average deviation. If we do the same for the other player, then whoever has the smaller average deviation can be said to be the more accurate bowler. Accuracy in this case equals ability to cluster woods closely around the jack.

In statistics, the smaller the average deviation the more closely the data are clustered around the mean. You should also realize that our test of bowling accuracy does not require that each player bowls the same number of woods and in statistics standard deviations can be compared irrespective of the number of scores in different groups. (Though it should be noted that the accuracy of our estimates of bowling accuracy will increase the greater the number of woods.)

Our bowling example modelled a statistic called the 'mean deviation'. The standard deviation itself is more complicated to calculate because it involves squaring deviations from the mean at one stage and unsquaring them again at another by taking a square root. If, however, you keep the bowling example in mind you will understand the principle which underlies it.

In terms of our examination scores the standard deviation for working-class pupils is 12.96 and for middle-class pupils is 17.4. This is what you could see in Figure 11, but is difficult to express in words. If we had much bigger samples of pupils with the same standard deviations their scores would make graphs like those in Figure 12.

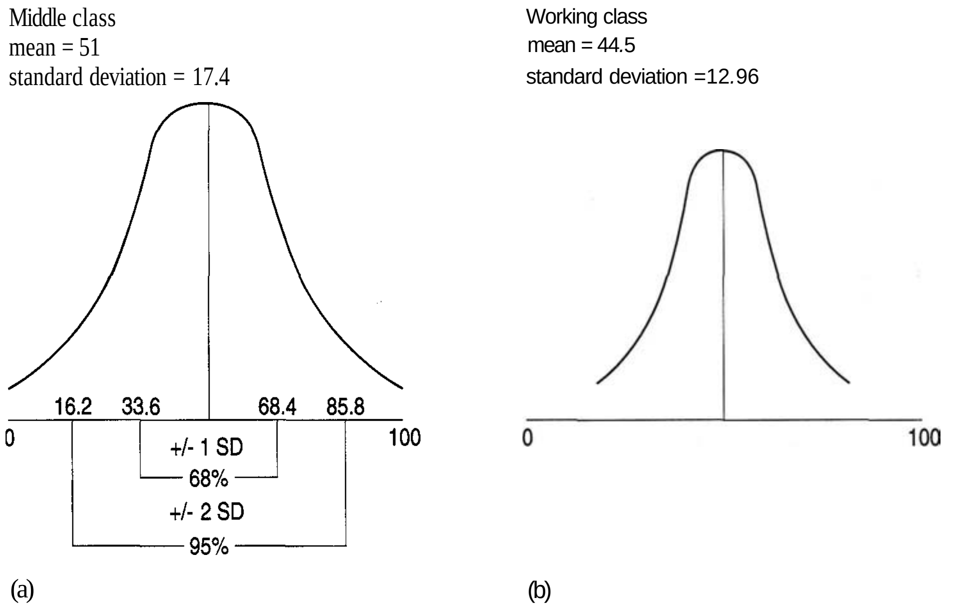


Figure 12
Distribution of scores (a) for middle-class pupils and (b) for working-class pupils

You will see from the annotations in Figure 12 that if we know the standard deviation of a set of figures we know what percentage of the scores will lie within so many standard deviations of the mean, although in this case the number of pupils in the class is too small for this to work out exactly. You should be familiar with this principle because it is exactly the same as that which underlay our card-dealing simulation. Unfortunately, this principle only applies when the distribution is symmetrical or 'normal', as shown in Figure 12. Ability and achievement tests are usually designed to give results which are normally distributed, as are national examinations such as GCSE and GCE A level. But much other data will not match this pattern.

We are not giving you the formula for working out standard deviations statistically because this is usually incorporated into the formulae for particular statistical tests, which you can find in standard statistical texts and which can be followed 'recipe-book style' without bothering too much about why the recipe is as it is. Furthermore, many of the more sophisticated pocket calculators have a function for standard deviations.

Scattergrams

Having introduced means and standard deviations, let us return to regression analysis. The first step is usually to draw a scattergram. The scattergram for our test scores and examination scores is shown in Figure 13.

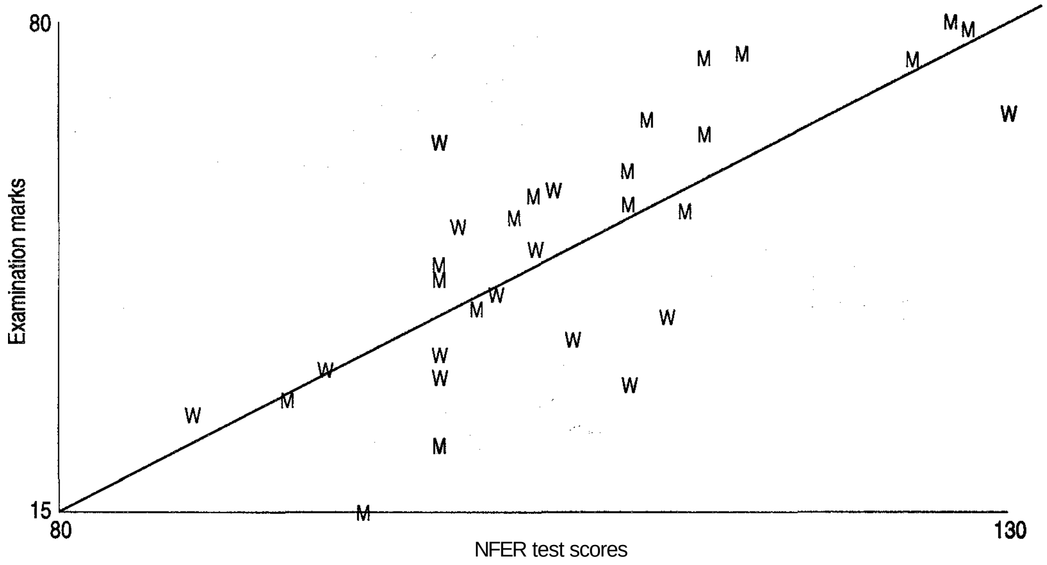


Figure 13 Scattergram for the whole data set

You will see that the bottom or *x* axis is the scale for the NFER test scores and the vertical or *y* axis is the scale for the examination results. It is conventional to put the dependent variable on the *y* axis and the independent variable on the *x* axis. In this case, since the examination results cannot have caused the test scores, the test scores must be the independent variable. You will see also that we have differentiated working-class and middle-class pupils on the scattergram as W or M.

The pattern shown in the scattergram is one of positive correlation because, in a rough and ready sort of way, pupils who score more highly on the NFER test are more likely to score highly in the examination. The scores can be visualized as distributed around a line from bottom left to top right.

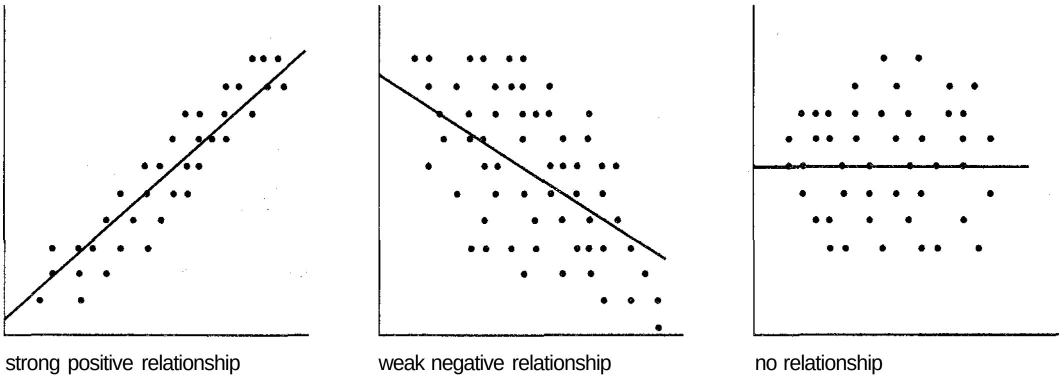


Figure 14 Scattergrams illustrating different relationships

The next step in regression analysis is called fitting a line, producing what is called the regression line. (In simple regression analysis this is always a straight line). We have already fitted the line to the scattergram. It represents a kind of moving average against which the deviations of the actual scores can be measured. If you remember our bowling example it is as if the bowlers had to place their bowls in relation to a moving jack.

The calculations for fitting the line take all the data and work out the average relationship between scores on the NFER tests and scores on the examination. The result is a statement, such as, on average *x* number of points on the *x* scale equate with *y* number of points on the *y* scale. This average is calculated from the performances of all pupils, not just those scoring 100 on the test. Suppose, for example, the result is that on average pupils scoring 100 on the NFER scale score 41.65 in the examination. Now we have a way of dividing pupils scoring 100 on the NFER scale into those who actually score above and those who actually score below this average.

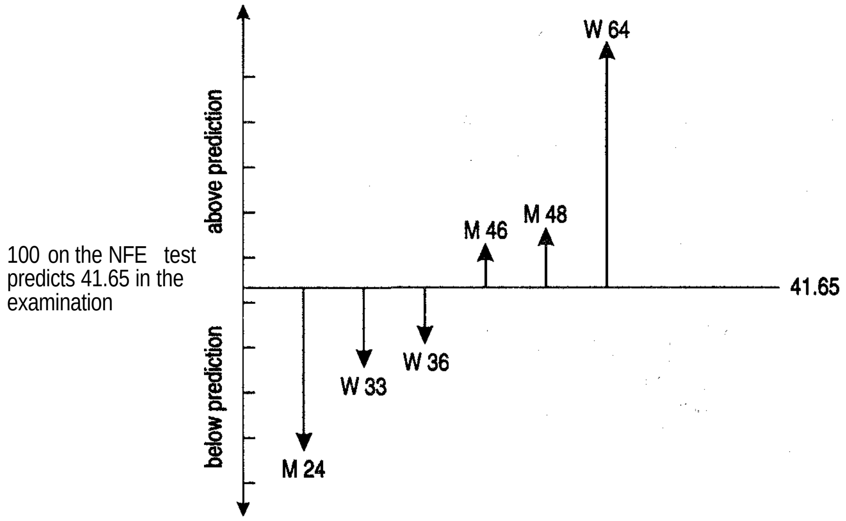


Figure 15 Identifying those pupils who scored 100 on the NFER test who score above and below the identified average

You can now see what is meant by the 'residuals' in the table of NFER and examination scores (Table 25). A residual is the figure produced by subtracting the actual score (on the examination here) from the score predicted (by the relationship, on average, between the NFER test score and the examination score). In principle, it is exactly the same as the result of subtracting expected figures from observed figures in a chi-squared test. Indeed, in a chi-squared test these are also called residuals. The regression line functions in the same way as the expected figures in a chi-squared test.

Remember that in this particular case we are interested in whether the gap between working-class pupils and middle-class pupils increases in the first three years of secondary school. If it does, we should expect to find more working-class pupils scoring below the regression line and more middle-class pupils scoring above. What we have done in effect is to control for ability (as measured by NFER test score) so that we can measure differences in examination achievement for pupils of the *same* ability from *different* social classes.

We could follow our bowling example and actually use the graph to measure this, physically measuring the distances above and below the line for each pupil of each social class and calculating a standard deviation for each social class of pupil. Physically measuring the residuals in this way, however, would be a very tedious procedure. Instead we can use any one of a number of statistical 'recipes' to reach the same result.

Demonstrating how to calculate a statistical regression goes beyond our objectives for this module. We shall, however, show you visually what this means. First, look at the original scattergram (Figure 13). You should be able to see that more middle-class pupils have scored more in the examination than might have been predicted from their NFER score and fewer have scored less. The reverse is true for working-class pupils. We have in the scattergram a visual display of a gap opening up between working-class and middle-class pupils, when previously measured ability is held constant.

Chance and significance

Any interpretation of our data as suggesting that social-class differences are involved stands or falls on the assumption that the differences between the examination scores of working-class and middle-class pupils are not simply due to the chance combination of particular pupils and particular happenings in this particular school class. Statistically speaking, there are no grounds for generalizing our findings for this school class to the ability band as a whole, the school as a whole, or to the whole age group in the country, because it is very unlikely that any single school class is statistically representative even of its own school and highly improbable that any school class is representative of the national age group. This is not a problem that is confined to quantitative research. It is one associated with the study of any naturally occurring group, such as a school class, whether by quantitative or qualitative methods. Had our data been drawn from a real school class we would have regarded the findings as interesting and worth following up by studying other school classes, but we would have had to have been circumspect about how far the findings could be generalized.

Chance factors might even undermine the validity of our findings as they relate to the pupils we are studying. That is to say, while it might appear that differences in examination scores are related to social class, they may instead be caused by chance. If you think of real examinations you will realize that many factors are each likely to have some small effect on examination performance: a cold, a broken love affair, a misread question, a cancelled bus, an absence from a critical lesson, the fortuitous viewing of a useful television documentary, the lucky choice of topics for revision - plus the vagaries of examination marking. Sometimes, of course, all these chance factors will cancel each other out, but occasionally they may fall, by chance, in such a way as to depress the performance of one particular group of pupils relative to another. Thus factors that are actually unrelated to social class might, by chance, add to the scores of one social class and subtract from the scores of another.

What we need is some way of estimating the likely play of chance factors that might skew the results in this way. This, of course, is the purpose of testing for statistical significance. The question in this case is whether or not the scores of the thirteen working-class pupils (and hence the eighteen middle-class pupils) are significantly different from those of the whole class. We might ask this question in relation to differences in terms of how well NFER test scores predicted examination scores, or more simply in terms of whether there is a statistically significant difference in examination scores between the two groups. The second logically precedes the first, since if there is no significant difference in examination scores between pupils from the two social classes then there is nothing to explain.

If we said that the examination scores are not significantly different it would mean that we could often draw random samples of thirteen pupils from the thirty-one, with scores departing from the profile of the whole class to the same extent as the scores of the thirteen working-class pupils differ from those of the whole class.

To put this into more concrete terms, imagine our class of pupils as a pack of thirty-one cards, each with an examination result written on it. Think of shuffling the pack and dealing out thirteen cards in order to produce two decks, of thirteen and eighteen respectively. Think of doing this again and again and again; each time recording the mean examination score of each pack. Intuitively you will know that the chance of dealing the thirteen lowest scoring cards into a single deck and the highest scoring eighteen cards into the other deck is very low. Similarly, so is the chance of dealing the thirteen highest scoring cards into a single deck very low. The vast majority of deals will result in something in between.

With this in mind, imagine asking someone else to divide the pack into a deck of thirteen and a deck of eighteen, by some unknown criterion. Again, intuitively, you know that if the mean score of one pack is very different from the mean score of the other it is much less likely that your collaborator used a random technique for the deal. In the same way, if the aggregate examination scores of our working-class pupils are very different from those of our middle-class pupils it is less likely that chance alone determined them and likely that there is some factor related to social class at play.

That is the logic of statistical testing. The question that remains is that of how much difference between the groups would allow us safely to conclude that our working-class pupils were a special group, rather than randomiy allocated.

Given the kind of data available to us here, which is at least of ordinal level, we would normally use more powerful statistical tests to establish this than a chi-squared test. A t-test would be a common choice, but since you are familiar with chi-squared tests, it is the one we shall use. To do this we shall degrade the data to convert it to the nominal level, by grouping pupils into two achievement categories: those scoring above the mean for the class and those scoring below the mean. We can then conduct a chi-squared test in the way that should now be familiar to you. For this operation the expected figures are those deriving from calculating what should happen if there were no difference in the distribution of working-class and middle-class pupils between the high- and low-scoring categories (that is, if high and low scores were distributed proportionately).

Table 27 Observed distribution of examination scores for working-class and middle-class pupils (0) compared with those expected if scores were distributed proportionately (E): pupils scoring above and below the school-class mean of 51.03

	Middle-class pupils		Working-class pupils		
	0	**E**	**0**	**E**	Total
above mean	12	9.3	4	6.7	16
below mean	6	8.7	9	6.3	15
Total	18		13		31

chi squared = 2.57, df = 1, $p< 0.2$ (20%)

From this calculation you will see that significance does not reach the 5% level and this means that if we kept on dealing out thirteen cards at random then we would quite often come up with results which departed from a proportional distribution of scores to this degree: in fact, about 20% of samples would be expected to deviate from a proportional distribution to this extent.

Inconclusive results

People are often very disappointed when they achieve inconclusive results and it is worth considering why the results are inconclusive in our case. On the one hand, there are the possibilities that the set of data contained too few cases to give a significant result, or that the particular school class chosen for research was odd in some way. These considerations suggest that we should expand our data set and try again, because so far we do not have enough evidence to accept or reject the idea that something about social class, other than ability, affects performance in secondary school. On the other hand, if we did investigate a larger set of data and still came up with the same kind of results, we should have to conclude that the results were not really 'inconclusive'. They would be conclusive in the sense that as we achieved the same findings from bigger and bigger sets of data, so we should become more confident that social class is not an important factor determining examination results, once prior ability is held constant. This uniikely conclusion would be a very important result theoretically and in terms of educational policy - it would run against the findings of many previous studies.

Activity 61 (allow 10 minutes)

Suppose well-conducted, large-scale research on educational achievement and social class showed that pupils aged 16+ and from different social backgrounds were performing, on average, much as might be predicted from tests of educational achievement at age 11+. What interpretations might you make of these findings?

One possible interpretation would be that secondary schooling had much the same average effect on all pupils irrespective of social class. In other words, the differences already present at age eleven persisted without much change to age 16+. In this case you would be interested in whether social-class differences in educational achievement were to be found in primary school, or whether differences in achievement were produced by factors that were uninfluenced by pupils' schooling at all, such as innate ability or parental support. How one would translate this kind of finding into policy terms would depend on whether you believed that secondary schools ought to level up (or level down) educational achievement between pupils of different social classes, in terms of predictions made at age eleven for performance at age sixteen.

Value-added studies and the school effect

In the past, it has been very difficult to assess the impact of different factors on a pupil's achievement at school because of an absence of relevant data. The education reforms of the late 1980s and early 1990s have changed the situation in England and Wales somewhat. The government decision that all secondary schools should publish their examination results, and the introduction of testing at the ages of seven, eleven and fourteen, promises to provide kinds of data not previously available. Of course, there has been much controversy surrounding these reforms and this has hinged in part on the extent to which the quality of a school can be judged by examination results at 16+ and 18+, without taking into consideration the different profiles of ability that each school intake represents.

Activity 62 (allow 30 minutes)

For a somewhat light-hearted treatment of this controversy read the piece called 'Publish and be damned. The problem of publishing examination results in two inner London schools' by John Gray (Article 5 in Reader 2).

Various alternatives to evaluation by raw results have been suggested, ranging from rather ambitious attempts to control for the social class and ethnic composition of schools, to rather less ambitious attempts to control oniy for prior educational achievement. Paradoxically, perhaps, the same government that has set its face against the publication of adjusted results has also been the government that has put in place the means to make such calculations routinely in the future. SATs tests should provide useful base-line data for judging the performance of pupils and, hence, the performance of their schools some time after the tests. SATs are allegedly criterion-referenced, with the same norms used for the entire national school population of England and Wales; and SATs scores provide data that are at least at the interval level. With the SATs data available, and assuming those data have a reasonable level of validity, very little research effort will be needed in the future to collect the information necessary for studies of 'the value added' to pupils over a period of time. To the present time, studies with this objective have been few and far between, and very expensive.

Activity 63 (allow 40 minutes)

In this activity you should read 'Beyond league tables. How modern statistical methods can give a truer picture of the effects of schools' by Ian Schagen (Article 6 in Reader 2). Read the entire article by all means, but we are principally interested in the section on 'Multilevel modelling' and not in the section on 'Data envelopment analysis'.

As you read the relevant section of Schagen's paper you will realize that it gives you another introduction to the principles behind regression analysis. In our example we used a very simple technique of regression analysis, comparing the relationship between only two variables - test score and examination result. It may not be apparent from Schagen's paper that the techniques he is describing involve very powerful statistical methods of multiple-regression analysis, which enable many variables to be inspected for co-variation simultaneously. As you read the paper we would like you to make sure that you understand the following terms:

- multi-level modelling and the term 'level' which it includes;
- fixed and random effects.

Read the relevant section of Schagen's article now.

Schagen's paper gave you a simple introduction to multi-level modelling. In a moment, we shall ask you to read a more complicated article which reports on a major multi-level study. This has data drawn from six LEAs and places the results in the context of the other multi-level studies that had been published by 1990. First, though, we need to outline the concepts of median and decile measurement, since this article relies on these.

The median and percentile measurement

As a measure of central tendency, the median is a useful device for setting up comparisons between groups, so long as a single scale of measurement is used. The median is a score which divides a group into two halves: one half with scores on or above the median and one half with scores on or below it. With an even number in a group the median lies between the two scores in the middle of the group. If you take the examination scores in our set of data (Table 25), there are thirty-one pupils and so pupil number sixteen has the median score, which is 53. Figure 16 shows this in another way. It is a cumulative frequency graph or 'ogive', where the horizontal axis shows the number of pupils scoring 'up to but no more than' and the vertical axis shows their scores.

On the graph we have drawn in the median with two co-ordinates. One runs to the 50% point (pupil number 16) and the other to the score of 53.

This principle can be extended to show what percentage of pupils achieved what scores. Sometimes this is done in terms of deciles, dividing the pupils into ten groups from the lowest scorer to the highest scorer. More common is the identification of quartiles. Just as the median divides the group into two halves, so the first quartile marks off the lowest scoring 25% and the third quartile marks off the highest scoring 25%. The median is of course also the 'second quartile' or the 'fifth decile'.

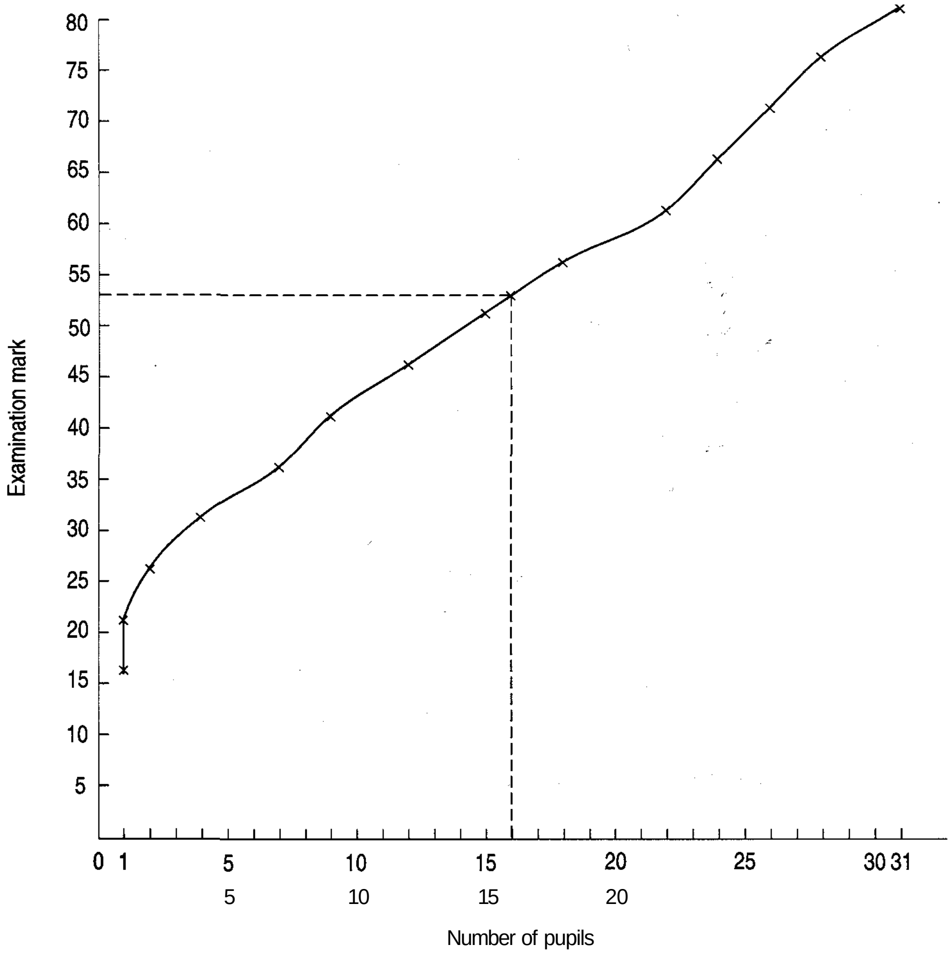


Figure 16 Ogive showing pupils' scoring at different levels in the examination

Activity 64 (allow 15 minutes)

On the ogive shown in Figure 16 draw in the co-ordinates for the first and third quartiles.

You will have found that the score for the first quartile is between 34 and 36, since 25% of 31 is 7.75 and the seventh from bottom pupil scored 34 and the eighth from bottom 36. For the third quartile the score is between 60 and *64.* There are ways of making a more exact calculation, but this is good enough for working on a graph.

The median alone gives a measure of the central tendency, but by itself does not show how variable the data are. For this, we need the interquartile range, which is calculated by subtracting the first quartile from the third. For our data, the interquartile range is 31. The median and the interquartile range clearly provide more information than the median alone. Even without a graphic display it is possible to visualize that there is more dispersion in the data where the median is 53 and the interquartile range is 31, than where the median is 53 and the interquartile range is 15.

To retain even more information, one can use percentiles. This sets up a kind of map on which the position of any particular pupil can be located. If schools want to do this they often just use a rank ordering technique: second out of 31, or fifteenth out of 20. Using percentiles you can say that a pupil scoring 70 is within the top 25% or above the third quartile. Do note, however, that this '25%' is 25% of pupils, not 25% of marks.

Activity 65 (allow 20 minutes)

You should now read the introduction to 'Estimating differences in the examination performances of secondary schools in six English LEAs: a multi-level approach to school effectiveness' by J. Gray, D. Jesson and N. Sime (Article 7 in Reader 2). Pay particular attention to the four points that appear at the end of the introduction. You might like to relate them to our previous discussion in terms of Figure 17.

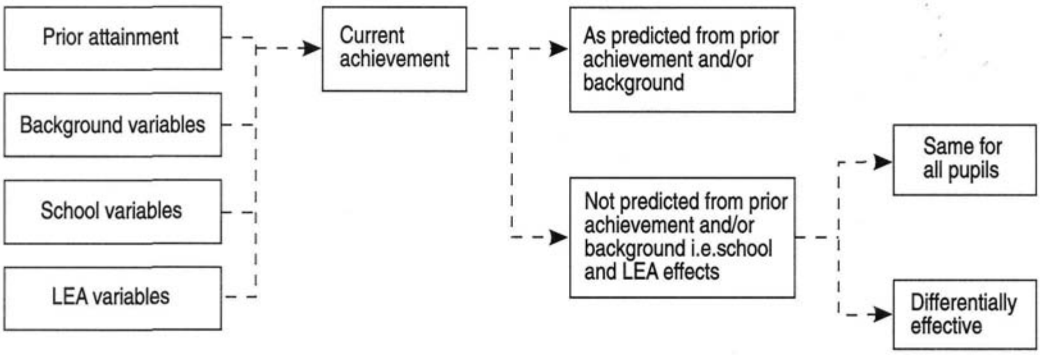

Prior attainment
Background variables
School variables
LEA variables
Current achievement
As predicted from prior achievement and/or background
Not predicted from prior achievement and/or background i.e.school and LEA effects
Same for all pupils
Differentially effective

Figure 17 The logic of school-effectiveness research

Activity 66 (allow 15 minutes)

Now read the section headed 'Sources of data' in Gray *et al.* in Reader 2.

It is worth noting that while a very large sample of pupils is involved, drawn from a large number of schools in six different LEAs, there is still the problem of how representative these pupils are of larger school populations such as that of England and Wales as a whole. Hence there is the problem of generalizing more widely from this study. Furthermore, note that different data are being collected from different schools, so like is not quite being compared with like.

Now you understand the median and the interquartile range, the box and whisker plots in the next section of Gray et *al.* should not give you too much difficulty, so long as you remember that the median is here a 'median mean'. That is, it is the median score of all the arithmetical averages of CSE/O-level results for the schools within an LEA in Figure 1 of the article.

Activity 67 (allow 10 minutes)

Now read the section headed 'The distribution of exam results across pupils and schools' in Gray *et al.* Why did they choose to compare schools and LEAs in terms of interquartile ranges (that is, the middle 50% of pupils)?

The choice to focus comparisons on the middle 50% is very sensible because the performance of a school varies greatly from year to year. In another study, Goldstein (1987) showed that schools considerably changed their position in the league table from year to year. Variations between schools or LEAs from year to year, however, are likely to be much less if only the middle scoring 50% of pupils are considered.

Activity 68 (allow 10 minutes)

Look at Table 2 in Gray *et al.* Here variation between schools within the same LEA is being shown. Each line represents a set of schools from a single LEA, although some LEAs appear twice with different collections of schools. The first column shows the amount of variation between schools in each set, without controlling for pupils' characteristics or pupils' prior achievements. In this sense it is equivalent to comparing schools on their 'raw' examination results. The second column 'fixes' the comparison by controlling for pupils' characteristics or prior achievements: that is, it compares and contrasts the performance, in different schools in the same LEA, of pupils with the same characteristics or the same prior achievements. Note that only for LEAs 5 and 6 were data on the pupils' prior achievements available. For the other LEAs all that is controlled for is the social background of the pupils. The third column expresses the variation between schools which is 'left-over' after controlling for the differences among pupils. This is the figure available for interpretation as 'the school effect'.

Activity 69 (allow 15 minutes)

Now read the sections headed 'The variances in pupil performance potentially attributable to schools' and 'Differences between similar pupils in different schools' in Gray *et al.* When you read the first part of this, jot down the estimates for the school effect claimed by different authors, Gray *et al.'s* own estimate and the figures in LEA 2.

You will see that there is a considerable consensus that the effects of school on a pupil's performance, independent of social characteristics and prior performance, are usually rather low: 'pupils attending more effective schools could be expected to obtain grade enhancements from say D to B in two subjects ... compared with their counterparts in less effective schools' (Gray *et al.*, 1993, p. 129).

On the other hand, the figures for LEA 2 show that sometimes the difference between the most and the least effective school can be much greater than this.

Activity 70 (allow 30 minutes)

Read the section headed 'The evidence for differential effectiveness' in Gray *et al.* The issue here is whether schools may serve some groups of pupils well, while serving others badly. Where the groups have been social classes, or genders, or ethnic groups this has often been the burning issue in educational research. You should understand the idea of 'differential slopes' if you remember the scattergrams discussed earlier. Note down for yourself what Gray *et sl.* conclude.

Activity 71 (allow 20 minutes)

In the next section of Gray *et al.* they review the kinds of input data which can be used in multi-level modelling.

Read the next ten paragraphs of Gray *et al.* which appear under the heading 'Explanatory power: substantive and statistical issues'. Draw on this reading and on your own ideas to fill in Table 28 below, ranking the data in terms of desirability for multi-level modelling (1 for data of first choice). Also note down the main advantages and disadvantages of each kind of data.

Table 28

	Rank	Advantages and disadvantages
Finely differentiated measures of prior achievement		
Grouped measures of prior achievement		
Tests of ability at or around the time when outcome measures are produced		
Aggregate measures of ability: e.g. percentage of pupils with below average reading scores		
Individual information on 'social background'		
Aggregate measures of social background of pupils in a school		
Neighbourhood characteristics		

In a moment we are going to ask you to read the remainder of the article by Gray *et al.*, but you might welcome some guidance for this. In the latter part of the section headed 'Explanatory power: substantive and statistical issues', Gray *et al.* draw attention again to the different estimates of school effects that arise from using different input data: measures of prior achievement as opposed to measures of social background. They do this by juxtaposing the estimates of the school effects that derive from LEA 6A, where measures of prior achievement were available, and from LEA 2A, where differences in social class were the input measurements. You may not find their diagrams (their Figure 5) very enlightening, so we have provided an alternative to the lower parts of their figure.

To read our diagrams (Figures 18 and 19), recall that in statistical terms 'accounted for by' or 'explained by' means 'predicted by'. Thus, in Figure 18, 56.3% means that if we knew the prior achievements of the pupils we could improve our prediction of the pattern taken by the results by 56.3%. If we knew which school they went to, we could improve our prediction by a further 2.2%.

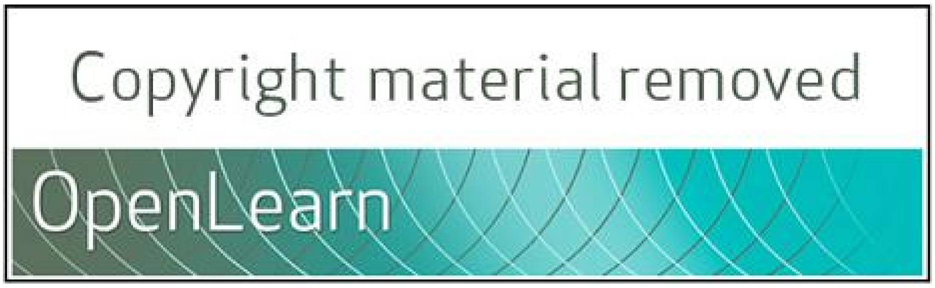


Figure 18 Variance in fifth-year examination results attributable to prior achievement and to the effects of schools in LEA 6A (derived from Gray et al., *Figure 5)*

Figure 19 Variance in fifth-year examination results attributable to social class and to the effects of schools in LEA 2A (derived from Gray et al., *1990, Figure 5)*

In addition, you may be confused by the authors' introduction of a second and smaller measure of school effects by comparison with what you read in their Tables 2 and 4. In their original estimates 100% was that variance which could not be explained with reference to prior achievement or social class (that is, residual variance), whereas in the later calculations 100% is all variance shown in the examination results (total variance). Since total variance is the bigger, school effects will constitute a smaller percentage of it. (Variance is a measure of dispersion calculated from adding all the squared deviations from the mean, and dividing by the number of scores).

Activity 72 (allow 20 minutes)

Using the notes above, now read the remainder of the paper by Gray *et al.* in Reader 2.

7.12 CONCLUSION

As things stand, the accumulated results from multi-level studies suggest that what different secondary schools do to their pupils and what one secondary school does differently to different pupils, are much less important factors in explaining differences in educational achievement than is the social background of the pupils. And they are even less important than the prior achievement of the pupils. The effects of primary schools remain a largely uninvestigated area so far as multi-level modelling is concerned, but what quantitative research exists leads to the same conclusion.

The research we have been discussing, particularly that reviewed by Gray *et al.* is, for all its flaws, currently the most representative picture of what is happening in the educational system in England and Wales so far as conventional measures of educational achievement are concerned. Similar research exists for Scotland, carried out by researchers working at the Centre for Educational Sociology at Edinburgh (Cuttance, 1988; Willms, 1987). Quantitative studies of this kind use large samples, which are either chosen to be representative or are chosen to span a range of different educational institutions. By contrast with these kinds of study, an ethnographic investigation of a school, which usually boils down to a detailed study of a few school classes or less, has no comparable claim to be representative. Generalizing from ethnographic (or indeed from experimental) studies to the educational system as a whole is extremely problematic. Small-scale studies may well provide inspiration as to what mechanisms actually create the causal links suggested by large-scale quantitative research, but without the latter we can never know how important observations at school level are.

At the same time, we must remember that large-scale quantitative studies buy representativeness at a cost. There are respects in which the validity of small-scale qualitative studies are likely to be greater. When reading the results of quantitative research and engaging in the sort of manipulation of figures in which we have been engaged here, it is easy to forget that the validity of conclusions based on quantitative data hinges on the extent to which the data actually measure accurately what we are interested in. Early on in this section we noted how the operationalization of concepts in quantitative research often raises questions about whether the phenomena of concern to us are being measured, or at least how accurately they are being measured. This problem is not absent in the case of qualitative research, but in the latter we are not forced to rely on what numerical data are already available or can be easily obtained from a large sample of cases. In assessing any study, then, we must always ask ourselves about the validity of the measurements involved.

In public discussions and policy making, there is often a tendency for these problems to be forgotten. Worse still, sometimes the use of quantitative data comes to structure our thinking about education in such a way that we start to believe, in effect, that the sole purpose of schools is, for example, to produce good examination results. In this way we may treat these as satisfactorily measuring academic achievement when they do not. We may thereby ignore other sorts of effect that we might hope schools would have on their pupils which cannot be measured so easily. Quantitative analysis provides a useful set of techniques, but like all techniques they can be misused.