

**ESRC Teaching and Learning Research Programme  
Research Capacity Building Network**

# **An introduction to the simple role of numbers in social science research**

**Stephen Gorard, Ken Prandy and Karen Roberts**

**Cardiff University School of Social Sciences**

**October 2002**

Occasional Paper Series  
Paper 53



**ESRC TLRP Research Capacity Building Network**  
Cardiff University School of Social Sciences  
Glamorgan Building  
King Edward VII Avenue  
Cardiff CF10 3WT  
[www.cardiff.ac.uk/socsi/capacity](http://www.cardiff.ac.uk/socsi/capacity)

Tel. 029 2087 5345  
Fax. 029 2087 4678



# **An introduction to the simple role of numbers in social science research**

Stephen Gorard, Ken Prandy and Karen Roberts

Cardiff University School of Social Sciences

[Gorard@cf.ac.uk](mailto:Gorard@cf.ac.uk)

This paper arose as part of our work for the ESRC-funded Teaching and Learning Research Programme Research Capacity Building Network (L139251106). Further information about the theme entitled ‘The simple role of numbers’ can be seen on our website [www.cf.ac.uk/socsi/capacity](http://www.cf.ac.uk/socsi/capacity), which also presents details of the other capacity building themes. Themes that may be of particular interest to readers of this paper include the role of interventions in educational research, complex statistical analysis, and combining data from different methods. These topics are not covered in this paper.

## **Introduction**

Not everyone understands why prices are as they are in shops, but everyone can check their change in a shop. A similar situation arises in research. Not all researchers will want, or be able, to conduct complex statistical analyses. Indeed, there is no need for them to be able to do so. But this is very different from an appreciation of the simple role of numbers in research. This paper is intended to show that confidence in dealing with numbers can be improved, simply by learning to think about them, and to express them more naturally. The paper also shows that a variety of important situations, relevant to educational research, are being routinely misunderstood (although the examples themselves are taken from a variety of fields, especially health research where there has been a stronger ‘quantitative’ tradition in the UK). The ignorance of false positives in diagnosis, the abuse known as the prosecutor fallacy, the politician’s error, missing comparators, pseudo-quantification and many other symptoms of societal innumeracy have real and tragic consequences. They are part of the reason why no citizen, and certainly no researcher, should be complacent enough to say ‘I don’t do numbers’.

The paper also shows that much of ‘statistics’, as traditionally taught, is of limited use in much current educational research. If research does not use a random sample then significance tests and the probabilities they generate are rather meaningless. It is, therefore, pointless for all researchers to learn a great deal about such statistics. This is especially so given that there is such general misunderstanding of the much simpler role of numbers in research. Papers are still being published using complex techniques, such as multi-level modelling, to overcome limited bias in their sample, but expressing changes over time using differences in percentage points (for example, see below). The latter is a more important problem, but it appears as though the use of a complex technique becomes a kind of magic bullet for some. Actually, most of the data we generate in our research is approximate and rough, and even the most elaborate statistical treatments cannot overcome this. Long term what we need, therefore, is to create better datasets preferably through active approaches such as experiments, and to ensure that our treatment of these datasets is first of all simple and robust. ‘One of the most tempting pitfalls... is to use over-sophisticated techniques’ largely because the computer makes it so easy (Wrigley 1976, p.4). The most complicated statistical approaches cannot overcome irrationality, lack of comparators, or pseudo-quantification. Everyone needs to know more about numbers to overcome problems of reporting (such as missing denominators, or unspecified averages). ‘I am not talking here about high level statistical skills... [but] concerned more with data and data quality than endless statistical tests’ (Plewis 1997, p.10). Perhaps it is time for some of those who use complex techniques somewhat mindlessly to re-assess the simple role of numbers, as well as time for those who refuse to use numbers at all to find out what they are missing. Education managers and practitioners deal with issues of risk assessment routinely, and sometimes, perhaps, unconsciously. It is certainly important that practitioners and policy-makers, who use the research evidence of others, have a better idea of its warrant (Schagen 2002).

With a good dataset, analysis is usually rather easy. For example, given data for a population there are no probabilistic calculations necessary, no tests of significance, confidence intervals, and so on. Similarly, running a simple experiment, for example, with two large randomly selected groups and one outcome measure means that all you need to do is compare the scores of one group with the scores of the other in order to produce an effect size. The key is to have the comparator to make the comparison with – an obvious point perhaps but one that appears to have eluded several writers in the field of

educational research. Without a comparison we must end up with an incomplete consideration of the plausible alternatives. The danger of this has been seen in a number of areas, such as disaster analyses. Component temperature analyses of the space shuttle flights in which problems occurred gave no clue to the cause of the Challenger disaster, but when scientists looked instead at flights in which no problems occurred the issue became clear (Dawes 2001).

### **Why we all need numbers?**

On the one hand, both of the claims above (the simplicity of using population data and comparison groups) mean that the role of numbers in research can be fairly basic. On the other hand, there is still so much misunderstanding of even this simple role of numbers that it seems a key place to start our co-operative venture of capacity building. This should not be taken to mean that only simple approaches should be used, or indeed that only numbers should be used. No one method, or type of method, is intrinsically superior to any other. Rather, different methods are differentially suitable for answering specific kinds of questions. Therefore we should not always rule out simple methods in favour of more complex ones, nor rule out working with numbers. If you are really trying to find out something then you will use all and any means available to you - and these will naturally involve some numeric information. Some people have suggested that there should be more statistical ('quantitative') studies in social science research because this form of evidence is intrinsically preferable and of higher quality than other forms. On the contrary, one reason to encourage a greater awareness of statistical techniques among all researchers is that quantitative work is sometimes poor, but largely unchecked. But even this is only one of many reasons why *all* researchers should learn something about techniques for research involving numbers.

### *So we won't get fooled again*

The first and most obvious point is that the process of research involves some consideration of previous work in the same field. All researchers read and use the research of others. Therefore we need to

develop what Brown and Dowling (1998) refer to as a 'mode of interrogation' for reading and using research results. If we do not have any understanding of research techniques involving numbers then we must either accept all such results without question, or ignore all such results. In practice, many commentators attempt to create a middle-way of accepting some results and rejecting others even though they do not understand how the results were derived. This usually means that results are accepted on the basis of ideology, or of whether they agree with what the commentator wants to believe. This is not a social scientific approach to research.

### *Context is everything*

Whatever your choice of primary method, there is a good chance that your research will involve numbers, at least at the outset. You may wish, for example, to document the experiences of the growing number of homeless people from ethnic minority backgrounds. Whatever approach you intend to use (participant observation, focus groups, anthropology, and so on) you should start from a quantitative basis. In order to direct your search you would use as much information as is available to you from the outset. You need to establish not only how many homeless people there are, but also where they are, how the socio-economic and ethnic patterns of these groups change over time and space, and so on. Such figures, termed 'secondary data', already exist, and therefore a preliminary analysis of them is the best place to start any study. Only then can you sensibly select a smaller group (a sample) for more detailed study. Existing figures, whatever their limitations, provide a context for any new study which is as important as the 'literature review' and the 'theoretical background'.

### *Some techniques are common to all research*

The choice and use of a sample, for example, is a common phenomenon in all kinds of research using many different approaches to data collection and analysis. It is not specific to what have traditionally been considered as quantitative designs. Similarly, once a discipline or field, like social science, is mature enough then some of its arguments can be converted into formal structures involving numbers whatever their original basis (Boudon 1974). This helps to reduce ambiguity, clarify reasoning and reveal errors.

### *A false dualism*

It is actually impossible to conduct purely 'qualitative' or purely 'quantitative' work. Both approaches rely heavily on each other. Patterns are quantitative in nature, while measurements are of qualities, for example (see below). We would be better off abandoning this false dualism, and its associated, and wasteful, paradigmatic strife (Gorard 2002).

### *Because it is easy*

Above all, it is important to realise that what is termed 'quantitative' research is generally very easy. Much analysis in social science involves nothing more complex than addition or multiplication - primary school arithmetic in fact. Even this, along with any more complex calculations, is generally conducted by a computer. There is no need for a paper and pencil. There is no need to practice any sums, or memorise anything. The important thing is to think about the numbers you use, and the social and psychological processes they are intended to represent.

### **Bread-and-butter issues**

Before continuing to look at some key issues in the simple use of numbers in social science research, the paper first deals with a few basic ideas about numbers and measurement.

### *What is a measurement?*

Everyone knows that measurement is about 'the assignment of numbers'. What we measure are quantities and they are, well, quantitative, and that means they must involve numbers.

The idea in the statement above underlies much of what is taught and believed in the social and behavioural sciences. It has its roots in the work of the natural sciences, and in most everyday dealings

with the material world. Unfortunately, it involves a basic misunderstanding of the idea of quantity and leads to widespread practices that can only be described as pseudo-quantification (Prandy 2002). The most notorious example of this is IQ and the claim that ‘intelligence is what IQ tests measure’. The problem, though, is not, as is usually argued, whether IQ tests measure intelligence, but whether they really measure anything at all, in the sense of establishing a *quantity*?

Fundamentally, a quantity is based on relations defined by a *quality*, in the sense of a qualitative observation. Take the simple example of measuring length. This is central to our understanding of the material world in which we exist. ‘Space’ and associated terms also serve as a metaphor for much of our conceptualisation of that world. For clarity’s sake, let us imagine away all we think we ‘know’ about length, distance and so on. We may, therefore, start with only the vaguest intuition, that we might have as a result of experience (our ‘theory’), that objects have a property of ‘longness’. If we collect together a set of objects – it is easier if these are relatively simple objects in ‘longness’ terms, like sticks (and this idea of ‘simple’ in relation to a putative quantity gives us a clue as to the relative intractability of the ‘objects’ of social science.) Our intuitive ‘theory’ suggests that ‘longness’ involves one object projecting beyond another when they are placed side-by-side. This is our qualitative observation; and it is important to note its crucial relation to theory.

If we then compare all of these objects pair-wise, in each case we can note which one projects beyond the other. Remember, we must not make any assumptions. ‘Longness’ is still only an idea. We cannot, for example, decide which is ‘shortest’, compare it with the next ‘shortest’ and so on. Of course, in reality we know that would work, but that knowledge depends on our prior understanding of length and an ability to make, at least crude, quantitative judgements. If we start (as we should) with the objects in a random order, then the result of all the pair-wise comparisons is an apparently complex set of inter-relations, showing which objects stick out beyond which other objects. However, it is possible to sort the objects in such a way that at one end is the one that projects beyond all the others, and at other the one that is projected beyond by all the others, while those in between have the property that they project beyond all those to one side and are projected beyond by all those to the other. Another useful way of looking at this is that objects close together in the ordering are, with respect to the ‘projecting beyond’ property, most alike to one another, while those further apart are most unlike.

So, starting from a qualitative comparison and an apparently complex set of relations we are able to produce a simple relation of order. We have established a quantity. Of course this is only a beginning. 'Longness' is only a quantity with respect to this particular set of objects and we cannot be sure yet that it would extend to others. Moving from this point to length involves additional complications, as does extending the concept to include distance. However, the procedure described so far is basic to the measurement process. Quantities as simple as this are found even in the natural sciences, where the Moh scale of the hardness of materials, based on the qualitative observation of scratching, is still used. It is true that the ordering of objects that is associated with a quantity has a parallel with the ordering of numbers. This can be very useful, because it is usually more convenient to work with numbers than with objects (dividing six by two is easier than sawing a six-foot length of log in half). However, quantities and number systems are two quite distinct structures; the utility derives from the fact of the *parallel* between the two, their isomorphism. However, it is the establishment of a quantity that assures the isomorphism, *not* the assignment of numbers; no amount of 'quantification' will, by itself, establish a quantity.

What are the lessons of this for measurement in social science? First, forget about numbers; with luck, they will come back in, but only in their function as a parallel. Second, concentrate on testable theory as the source of the qualitative observations that are at the heart of quantities. Third, look at the relations between objects that are created by these qualitative observations. A major problem in social science is that, unlike sticks, human beings and their social creations are extremely complex objects. Answering yes to a question, being a member of a group, engaging in an activity and so on may well be elements of a putative quantity, but they are also likely to be influenced by other individual characteristics. As a result, these observations are subject to considerable variation or 'error'. Nor is it usually possible, as with the natural world, to create objects like the standard metre that enable quantification to go far beyond the simple ordering that we established. The fact that the objects of natural and social science differ somewhat is an argument for being sceptical about attempts by the latter to ape the form of quantification found in the former, but this is in no way the same as an argument for rejecting quantitative thinking altogether.

*What kind of measurements are there?*

The discussion continues by drawing a tentative but useful distinction between two types of numeric data. Although too much is often made of fine distinctions between numbering scales or levels of measurement, the novice analyst must learn to recognise the differences between descriptions of categorical information and real numbers.

'Real' numbers are those that it makes sense to do arithmetic with. So a simple test of identification would be - does it make sense for me to add or subtract these numbers? The number of years a steelworker has been employed in a factory is a real number. To find the difference in experience between two steelworkers we could subtract two numbers of years and find how many years more one steelworker had been at the factory. We can do this because the scale we use to measure time with has equal intervals all the way along. The difference between 99 years and 100 years is the same as that between 1 and 2 years, for example. A year is a year wherever on the scale we look. In this respect our quantification is rather like that of the natural sciences.

Categorical information, on the other hand, relates to categories only, and individual cases therefore cannot be subject to arithmetic operations. The sex of a doctor is a category, and we cannot subtract a maleness of one doctor from the femaleness of another to find their difference in gender. This restriction applies even where the categories are expressed as numbers. Whereas the length of my foot is a real number, my shoe size is a category (shoe sizes are not equal interval as childrens' sizes increase in smaller stages than adults'). We could add two lengths but not two shoe sizes. Arithmetic operations can, however, be conducted using the frequencies of categorical data. We could for example find a difference by subtraction between the number of male and female steelworkers in a factory, or find the total number of people with either of two shoe sizes. In fact, most social scientific data has elements of both types expressed as the number of things of a certain category.

Other authors give much greater attention to measurement theory and the issue of scales (see for example Siegel 1956). However, the first and clearest distinction is the one just introduced between

numbers we can add together, and numbers used to label categories or types of things. On reading a traditional statistical textbook you will be introduced early on to measuring scales called 'ratio', 'interval', 'ordinal' and 'nominal'. But both ratio and interval measures are real numbers and the difference between them will not make any practical difference. There are very few purely interval measures, and the kinds of statistical procedures you would use, at least for the beginner, are identical to those for ratio measures anyway. Both ordinal and nominal scales are categorical in nature, but in many practical situations ordinal values are treated in the same way as interval data.

### *Proportionate reasoning*

A local paper recently ran a front page story claiming that Cardiff was the worst area in Wales for unpaid television licences - it had 'topped the league of shame for the second year running'. The evidence for this proposition was that there were more people in Cardiff caught using TV without a licence than in any other 'area' of Wales (and it is important for readers to know that Cardiff is the largest city in Wales). Not surprisingly the next worst area in the league of shame was Swansea (the second city of Wales), followed by Newport, and so on. Everyone who has heard this story laughs at the absurdity of the claim, and points out that the claim would have to be proportionate to the population of each area. Cardiff may then still be the worst, but at present we would have to assume that, as the most populous unitary authority in Wales, Cardiff would tend to have the most of any raw-score indicator (including, presumably, the number of people using TV *with* a licence). Why does this matter? It matters because very similar propositions are made routinely in social science research, and rather than being sifted out in peer review, they are publicised and often feted. Examples include claims about the 'underachievement' of ethnic groups, and differences in attainment over time and between regions (indeed all of the problems discussed in this paper are relevant to UK educational research, see Gorard 2001).

These errors and misunderstandings are crucial, because the risks and uncertain gains we work with in practical educational research are probabilistic and are usually expressed as percentages or proportions. As the newspaper example reveals, this is the correct approach. Yet it is also surprisingly difficult to do properly (see below). What we hope to do in this paper is to suggest a new way of explaining, teaching

and communicating the calculation of probabilities with minimal recourse to percentages. Insight into complex numeric situations can be encouraged simply by taking more care in the presentation of probabilities. Almost anyone, however inexperienced, can calculate conditional probabilities of the kind that would even confound some experienced mathematicians. To make this possible, we mostly need to change the way we think about and represent probabilities, rather than simply improve our own computational ability.

As an analogy, consider this computational problem. In a standard knockout competition, such as a singles tennis tournament, if there are four players then there will be three matches in total – two first-round matches and a final. If there are eight players there will be seven matches in total – four first round matches, two second round matches and a final. Sixteen players leads to fifteen matches, and so on. It seems that for any number ( $n$ ), which is a power of two, there will be  $n-1$  matches in total. But if  $n$  is *not* a power of two, then unmatched players will have a bye in the first round. How many matches would be played in total for a tournament of 43 players? How could you prove your answer?

Before you start working this out with paper and pencil, consider the following. Each match has two players and only one of them goes through. So each match eliminates one player. To have a winner we need to eliminate all but one player. Therefore, however many players there are, the tournament needs  $n-1$  matches (42 matches for 43 players). This is a very simple proof of the general solution, and most people recognise it, but in practice even professional mathematicians struggle to find it for themselves (Dawes 2001). This is because the initial problem is phrased in such a complex way that many readers start looking for a complex answer. Once the problem is phrased more simply, the solution is obvious. We know that the social world is complex (and appears even more complex when closely examined). That, in itself, is not an interesting discovery; nor does represent an increase in useful knowledge. Let us take that for granted. The lesson from the knockout analogy is that, similarly, when dealing with difficult issues in our research we should try and represent them as simply as possible. This, in itself, is not easy. It takes hard mental work. But once done the simpler representation is easier to work with, easier to communicate, and easier to teach.

## **We don't always know what we think we know**

In order to back up some of the claims made so far, we present here a number of areas which suggest the need for a greater critical awareness of the role of numbers in research. Many of them are also counter-intuitive.

### *Conditional probability 1*

Imagine being faced with the following realistic problem. Around 1% of children have a particular specific educational need. If a child has that need, then they have a 90% probability of obtaining a positive result from a diagnostic test. Those without that specific need have only a 10% probability of obtaining a positive result from the diagnostic test. If all children are tested, and a child you know has just obtained a positive result from the test, then what is the probability that they have that specific need? Faced with problems such as these, most people are unable to calculate a valid estimate of the risk. This inability applies to relevant professionals such as physicians, teachers and counsellors, as well as researchers (Gigerenzer 2002). Yet such a calculation is fundamental to the assessment of risk/gain in a multiplicity of real-life situations. Many people who do offer a solution claim that the probability is around 90%, and the most common justification for this is that the test is '90% accurate'. These people have confused the conditional probability of someone having the need given a positive test result with the conditional probability of a positive test given that someone has the need. The two values are completely different.

Looked at another way, of 1,000 children chosen at random, on average 10 will have this specific educational need (1%). Of these 10 children with the need, around 9 will obtain a positive result in a diagnostic test (90%). Of the 990 without the need, around 99 will also obtain a positive test result (10%). If all 1,000 children are tested, and a child you know is one of the 108 obtaining a positive result, what is the probability that they have this need? This is the same problem, with the same information as above. But by expressing it in frequencies for an imaginary 1,000 children we find that much of the computation has already been done for us. Many more people will now be able to see that

the probability of having the need given a positive test result is nothing like 90%. Rather, it is 9/108 or around 8%. Re-expressing the problem has not, presumably, changed the computational ability of readers, but has, we hope, changed the capability of many readers to see the solution, and the need to take the base rate (or comparator) into account.

The same approach of simplification can also help us to overcome what has been termed the ‘prosecutor fallacy’. In judicial proceedings (and media reporting), forensic evidence (such as a fingerprint or DNA profile) is used to make a match with a suspect. Prosecutors tend to use the probability of such a match (e.g. 1 in 10,000) as though it were the reverse of a probability of guilt (9,999 in 10,000). However, they have to argue also that there is no human error in the matching process, that the match signifies presence of the suspect at the crime scene, that presence at the scene necessarily entails guilt, and so on. Above all, they have to demonstrate that the number of potential suspects is so small that a 1 in 10,000 chance is the equivalent of ‘beyond reasonable doubt’. If the crime took place in a city of 1 million people, and if we make the favourable assumption that potential suspects are limited to residents only, then 1/10000 means that 100 residents will have just such a forensic match. Thus, the suspect actually has a 1/100 probability of guilt (on this evidence alone). This is much higher than for an average resident of that city (and therefore germane to the case without being conclusive) but much lower than 9999/10000. The importance of this error, and others like them, is hard to overestimate in law, medicine and beyond. The same error occurs regularly in statistical testing in educational research, where researchers treat the inverse of the probability of their null hypothesis (see below) as identical to the probability of their conclusions. The gravity of this error would be hard to overestimate. But again, presenting the probabilities as frequencies makes the calculation much easier to follow.

### *Conditional probability 2*

Imagine a test to predict the sex of unborn children. Is it possible that the test is more accurate at predicting boys than girls, but at the same time that predictions of girls are more likely to be correct? Assume, for example, that the actual probability of boys and girls are equal, and that the predictor test is better than chance for both sexes. Assume also that the test is better for correctly deciding on a boy if

the child actually is a boy, than it is for correctly deciding on a girl if the child actually is a girl. Suppose it is 80% accurate for boys, but only 70% for girls, then there will be 20% misclassified boys and 30% misclassified girls. For simplicity the figures in Table 1 are presented for a base figure of 100 cases.

**Table 1 – Predicting sex of unborn child (from Dawes 2001)**

	Actual boy	Actual girl
Predicted boy	80	30
Predicted girl	20	70
Total	100	100

The table shows that 110 out of 200 cases will be predicted as boys (of whom 80 *are* boys) and only 90 cases as girls (of whom 70 *are* girls). Therefore, if the test predicts a boy then the probability that this is correct is  $80/(110)$  which is equal to 0.73 or 73%. If the test predicts a girl then the probability that this is correct is  $70/(90)$  or 78%. Thus, predictions of girls are more accurate than predictions of boys. Again, one point in this example is that expressing the problem in simple frequencies makes it easier to follow and to compute the relevant odds. The main point, however, is that the computation has to be done because the discriminatory power of this test is counter-intuitive. We need to think carefully about statements involving even simple numbers, because of the seductive nature of what we think we know about the nature of the ‘unexamined’ world.

### *Simpson’s paradox*

This is made even clearer in the following example. Imagine measuring the frequency of a characteristic, and finding that it is more prevalent in one group of people than another. Imagine then dividing the two groups by sex, and finding that for both sexes this characteristic is more prevalent in the opposite group to when they were combined. Is this possible? Put another way, could men in group A have more of this characteristic than in group B, and women in group A have more of this characteristic than in group B, while group B overall has more of this characteristic than group A? The answer is ‘yes’.

Suppose an education authority monitors 2000 students in terms of the way they are taught for a particular school subject. Of these, 1000 are given an experimental ‘treatment’, perhaps a new method of mentoring or curriculum material. The other 1000 are taught using the previous traditional methods for this subject. In a post-test, 600 of Group A (experimental) pass a particular test of achievement in this subject, while only 500 of Group B (control) pass. If the design of the experiment is adequate, then this is *prime facie* evidence that the new treatment leads to better results (Table 2).

**Table 2 – Pass rates of experimental and control groups**

Group	A	B	Overall
Pass	600/1000 (60%)	500/1000 (50%)	1100/2000 (55%)

Suppose the researcher then considers pass rates in terms of student background characteristics such as sex (Table 3). The observed pass rate was higher for female cases (67%) than for male (43%). But when disaggregated by sex and group the results lead to a kind of 'paradox' (attributed to Simpson 1951, but probably much older). The pass rate for Group A (60%) is higher than for Group B (50%), yet females in Group B (75%) do better than females in Group A (65%) while males in Group B (44%) also do better than males in Group A (40%).

**Table 3 - Pass rates by sex of patient and experimental group**

Group	A	B	Overall
Female	520/800 (65%)	150/200 (75%)	670/1000 (67%)
Male	80/200 (40%)	350/800 (44%)	430/1000 (43%)
Overall pass	600/1000 (60%)	500/1000 (50%)	1100/2000 (55%)

This seems impossible. It is one reason why investigations at one level of detail (aggregation) can only produce automatically safe results at the same level. Put another way, results should not be mindlessly aggregated up (or especially down). Just as an investigation at the school level, for example, may produce misleading results for the individual level, so an investigation at the individual level *could* produce quite different results at the school level. The safest course is to use the unit of sampling (whether school or individual, for example) as the unit of analysis *and* the level at which conclusions are

warranted. What this means is that the prior selection of the appropriate unit of analysis is a key step in the research process, and involves careful consideration of the relationship between any sub-groups and the subject of the investigation.

### *Ill-conditioning*

Is it possible that a tiny error in measurement, of the kind that routinely occurs in social science, can lead to a totally misleading result? For example, if we take a measurement containing an error of one part in one thousand, and then do some analysis with this figure, could our result have an error larger than the result itself?

We have to assume that any measurements we make contain some error. These could be caused by inaccuracy in measurement, corruption of stored data, transcription (copying) errors, or restricting our working to a certain number of decimal places or significant figures. To a large extent we behave as though the error component in our analysis remains constant, so that if we start with figures at a certain level of accuracy, we will end up with results at approximately the same level of accuracy. In some cases, known in extreme form as 'ill-conditioned' problems, this assumption is far from justified. Unless we track the potential propagation of these errors it is possible for our answer effectively to 'cancel out' the estimates we started with, and so contain a much larger proportion of error component than we started with. Consider the simultaneous equations:

$$400 = 200x + 200y$$

$$201 = 101x + 100y$$

Their solution is that  $x=1$  and  $y=1$ . If a measurement in the second equation is incorrect by less than half of one percent, then the true value of the first figure could be 200 (rather than 201), making the equations:

$$400 = 200x + 200y$$

$$200 = 101x + 100y$$

The solution now is that  $x=0$  and  $y=2$ . This is a totally different solution 'caused' by a small proportionate error in one term. Imagine the practical implications if  $x$  and  $y$  were the hypothesised components of an effective school, or a successful lesson. For some problems the introduction of an error component makes a large difference, and for some problems the error makes *all* of the difference. Certainly, the size and propagation of measurement error is a problem of a far greater importance for research than issues such as sampling variation, heteroskedasticity, or auto-correlation that are given far more attention in traditional methodological treatises. Most of the latter issues only reduce the 'power' of the analysis and can be remedied simply by using a larger sample. But if the measurements we make are inaccurate (or worse, meaningless) then no amount of statistical jiggery-pokery will lead to safe conclusions by itself. And the measurements we make in social science research are nearly always inaccurate.

#### *The politician's error*

Imagine a country of 100 million adults, of whom 50 million are male and 50 million are female. There are 1000 members of parliament (MPs or elected representatives), and all of these are male. The employed workforce is 50 million of whom 25.5 million are male. No great analytical skill is required to see that this imaginary country has a considerable political bias towards males. Similarly, it is easy to see that the country also has a slight employment bias towards males but that the political bias is much greater than the employment bias. None of the female half of the population are MPs, while 49% of women are in employment. Of the male population 0.001% are MPs, and 51% are in employment. We repeat, because of the importance of this point, that the ratio of male to female MPs is 1000:0 (equivalent to an infinity) whereas the ratio of male to female employed is 25.5:24.5 (equivalent to 1.04). Therefore the inequity among MPs is far greater than among the general employed workforce. Why emphasise this point? Because the most common 'method' used to analyse such data comes to the opposite, and wrong, conclusion. This purported method is used very widely in areas of social science research, in the media, and most frighteningly of all in policy documents and policy-making. It is the method of differences between percentages.

The argument goes like this. The percentage of male MPs is 0.001% and the percentage of female MPs is 0%, so the difference between them is 0.001%. The percentage of males in employment is 51% and the percentage of females is 49%, so the difference between them is 2%. Since 2% is much larger than 0.001% the lack of equity in general employment is greater than among MPs. This is a weak argument making several related arithmetic mistakes, yet many readers will have accepted this kind of 'analysis' at face value on many previous occasions. This is precisely the kind of example that leads us to argue that all researchers, indeed all good citizens, require some knowledge of what are termed quantitative research skills. So we won't get fooled again. Perhaps you do not believe that people get away with it. Consider another imaginary example, this time written as the start of a newspaper story.

***Girls leave boys in their trail!***

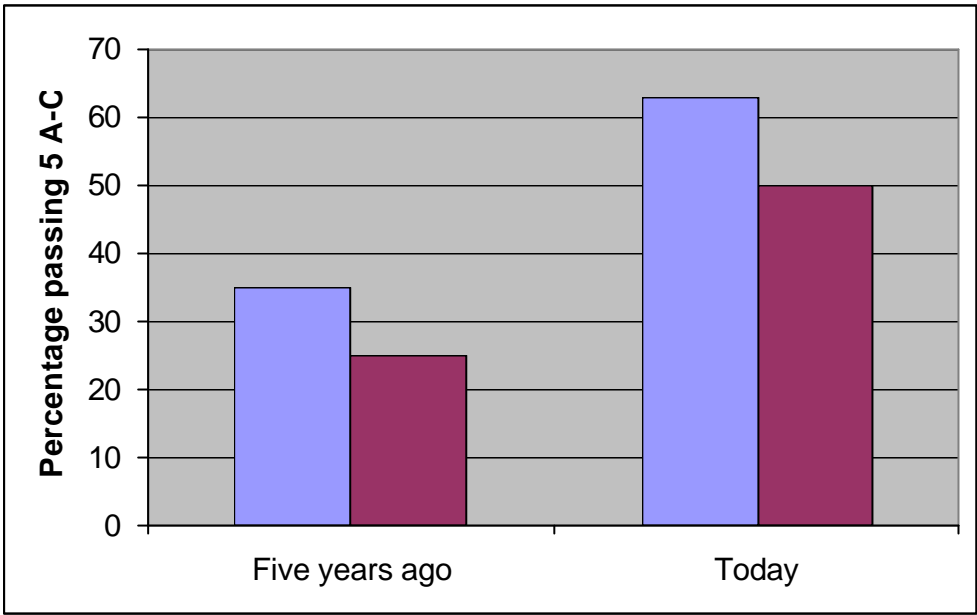
*The new GCSE results for England and Wales have just been released and they do not make pretty reading for the families of boys. While general levels of qualification continue to rise, the difference between the performance of girls and boys is growing to crisis proportions. Five years ago 35% of girls obtained the government benchmark of five good GCSE passes while only 25% of boys did. This year 63% of girls got five good GCSEs and only 50% of boys did. The gender gap has grown from 10% five years ago to 13% this year, reflecting the increasing problem of boys' underachievement which faces the education system. In fact, the minister for schools was quoted last night as saying that the growing underachievement of boys at school was one of the most serious problems faced by our society today....*

Such stories, using precisely these type of figures, are commonplace in the media (see Gorard et al. 1999 for a fuller list of examples). The logic is the same as in the example of the MPs above. In order to decide what is happening we cannot simply subtract two sets of percentages and compare the results. One of the main reasons for this is that the difference between two percentages is not itself a percentage. In the newspaper example girls are not doing 13% better than boys this year; rather they scored 13 percentage points higher than boys. The distinction is crucial. If we look at the figures as ratios, as we did for the MPs, we see that the proportion of girls to boys with five good GCSE passes five years ago was 35:25 (equivalent to 1.4, or a 40% gap in favour of girls). This year the proportion

was 63:50 (equivalent to 1.26, or a 26% gap in favour of girls). What the newspaper figures actually show is that the proportionate gap between girls and boys has fallen over time. Put another way, the scores for boys have doubled over five years (100% increase), while the scores for girls have only increased by 80%.

Of course, part of what is seductive about the percentage difference approach is that one can apparently see the gap changing over time on a graph. In Figure 1 the distance between the two bars is greater for the current score than for the previous score. This approach is used quite widely in some respected research reports, books and journal articles. Of course, an equivalent graph for our hypothetical example of elected and employed men and women (0.001 to 0, and 51 to 49) would show an even more extreme difference in distance, but still signifying nothing. Since all of the numbers change in size from one case to another, the question is not whether any percentage point difference has grown but whether it has grown more or less than the numbers between which it is the difference. Or, put more elegantly, 'the drawback with using the absolute difference in proportions to evaluate social reforms, however, is that the measure is largely driven by changes in the overall totals' (Heath 2000, p. 318).

**Figure 1 - The 'growing' gap between girls and boys**



Dawes (2001) makes a similar complaint concerning the use of symptoms in medical diagnosis. Imagine an illness that occurs in 20% of the population, and has two frequent symptoms. Symptom A occurs in 18% of the cases with this disease, and in 2% of cases without the disease. Symptom B occurs in 58% of the cases with the disease, and in 22% of cases otherwise. Which symptom is the better predictor? Many practitioners would argue that symptom B is the more useful as it is more ‘typical’ of the disease. There is a 16% gap (18-2) between having and not having the disease with symptom A, whereas the gap is 36% (58-22) with symptom B. Symptom B, they will conclude, is the better predictor. But while it seems counter-intuitive to say so, this analysis is quite wrong because it ignores the base rate of the actual frequency of the disease in the population.

In a group of 1,000 people, on average 200 people would have the disease and 800 would not. Of the 200 with the disease, 36 (18%) would have symptom A and 116 (58%) symptom B. Of the 800 without the disease, 16 (2%) would have symptom A, while 176 (22%) would have symptom B. Thus, if we take a person at random from the 1,000 then someone with symptom A is 2.25 times as likely to have the disease as not (36/16), whereas someone with symptom B is only 0.66 times as likely to have the disease as not (116/176). Put another way, someone with symptom A is more likely to have the disease than not. Someone with symptom B, on the other hand, is most likely *not* to have the disease.

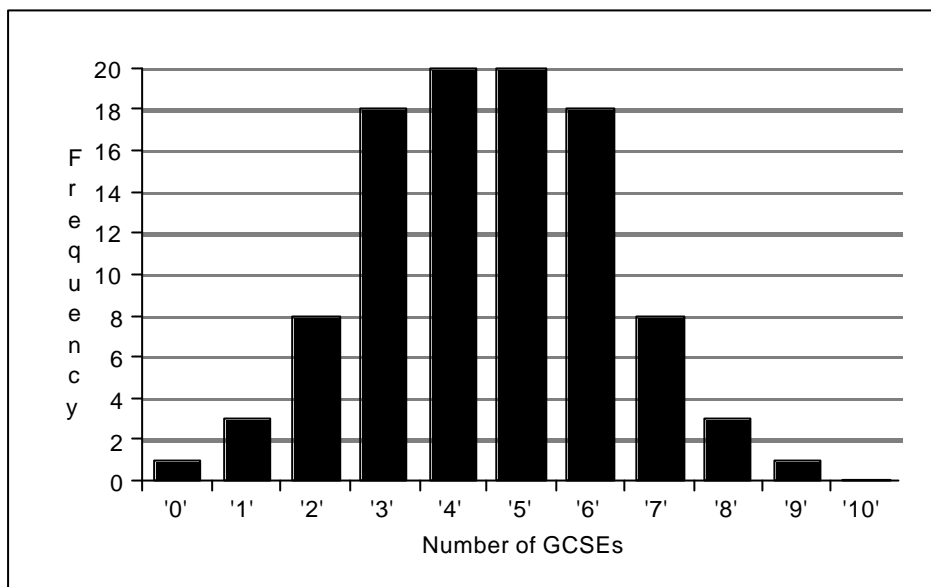
What we need for diagnosis are discriminators, rather than typical, symptoms. The more general conclusion is therefore the same as in the example of MPs, and of boys and girls. Simple differences between percentages give misleading, and potentially extremely harmful, results.

### *The importance of distribution*

Another reason why simple comparisons, such as those in the last section, do not work is that the figures they are based on may not have a straight line relationship with their underlying variables (Fleiss 1973). The percentage points used in the examples above are not real numbers (see above). To be real numbers, the interval between 10% and 20% (10 points) would have to equal the interval between 20% and 30% (10 points), for example. This is not always so, so the kind of arithmetic used to create the politician's error is wrong. Rather than being a straight line many patterns, trends and relationships in social science follow a traditional S shape. This consists of a threshold, below which any change in the x-axis produces little or no change in the corresponding value for the y-axis, then a line, where changes in x are linked to changes in y, and finally saturation, above which any change in the x-axis again produces little or no change in the corresponding value for the y-axis. A practical example, of saturation, is a difference of 50 percentage points between 40% and 90%. If the lower figure grows to 60% then the difference between it and the higher figure cannot be 50 points any longer, however much the higher figure grows. 100% intervenes as a limiting factor.

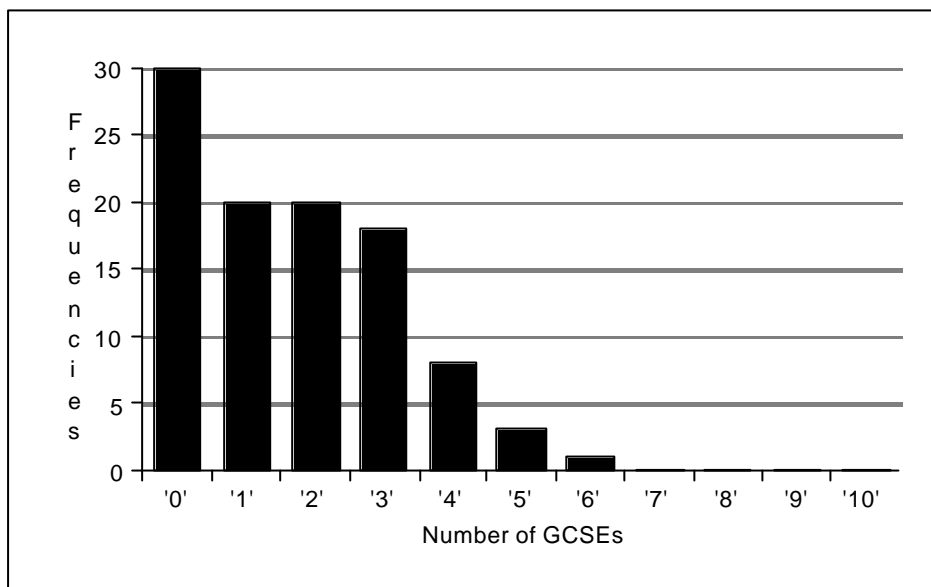
A similar logic applies at the other end of the percentage scale. The frequency of any population characteristic, such as the number of GCSEs per student, in any sizeable group is likely to be approximately normally distributed (i.e. to follow the traditional bell-shape). In Figure 2 representing a hypothetical school, exactly 50 of 100 students gain five or more GCSEs. If all of these students had gained one more GCSE then every bar on the graph would shift one place to the right, and the benchmark figure for the school would rise to 70%.

Figure 2 - Distribution of GCSEs among candidates (high score)



In Figure 3 representing a lower attaining school, only 4 of 100 students gain five of more GCSEs. However, the actual pattern of frequencies for Figure 3 is the same as Figure 2 but moved three places towards the origin of the x axis and therefore ‘squashed’ up by the limit of zero. In this case, even if all students gained one more GCSE each, then the benchmark figure for the school would only rise to 12%. Increases in terms of absolute percentage scores are clearly much more difficult for low-attaining schools (but this phenomenon is never acknowledged by policy-makers or school ‘improvers’).

Figure 3 - Distribution of GCSEs among candidates (low score)



These problems give the GCSE benchmark score the typical S shape rather than a straight line growth. Many other social science phenomena are similar in having the 'threshold' and 'saturation' phase created by the limits of 0 and 100%, which means that they must be handled with care. At the 50% mark, where the distribution is taller, a small movement along the x-axis (representing a change in the number of GCSEs per student, for example) would produce a disproportionately large change in the percentage attaining the benchmark. At either end (near 0 and 100%), a much larger change on the x-axis would be needed to produce the same effect. A more technical way of summarising this whole section is that 'method' of differences between percentages is not 'composition invariant'.

### **Composition invariance**

Consider the following example. A researcher collects secondary data from a government department about the number of students from each ethnic group at universities in the UK over the last five years. The researcher is also given access to the total number of students attending universities over the past five years. The intention of the research is to decide whether universities have widened participation rates for students from ethnic groups other than those classified as 'white'. The researcher calculates the percentage of the total student population who are in each ethnic group for each of the five years. The

researcher finds that the percentages at university of all ethnic groups other than white have increased, while the percentage classed as white has declined in each of the five years. These findings are published as stating that participation by ethnic minorities has widened, such that the student population is now a better reflection of the total population of the UK. But this conclusion is unsafe.

The researcher in the example above should actually be tracking two trends over time. The first is the proportion of ethnic groups in universities, and the second is the proportion of the same groups in the total population. It is, of course, possible for the proportion of ethnic minority students at university to increase but for the proportion of ethnic minority members of the population to increase even faster, meaning that universities are actually increasing their over-representation of the white population. What is needed is a way of comparing the number of ethnic minority students ( $e$ ), the number of ethnic minority members of the population ( $E$ ), the total number of students ( $t$ ) and the total number of people in the population ( $T$ ). One formula (for the segregation ratio, see Gorard and Fitz 2000) is:  $SR = (e/t) / (E/T)$ . If  $SR$  equals one then the group in question is perfectly represented. If it is less than one the group is under-represented, so that 0.5 means that the group has only half the expected representation. If the ratio is greater than one the group is over-represented, and so on. Once you have grasped this useful proportionate approach you can see that you could substitute males and females (or any categories you want) for white and ethnic minority. You could substitute owning a house or requiring heart surgery (or again any measure you want) for attendance at university. In practice, you may prefer to use the logarithm of the ratio, so that divergence from one becomes an equal-interval value. The segregation ratio is a general measure of unequal representation, which can be used to make safer comparisons over time, place and other categories because it takes into account changes in both of the proportions involved.

In our example, the segregation ratio has a key problem however, which is that it can only tell us about the overall figures. It might be the case that the number of white and ethnic minority (or male/female or whatever) students was a perfect reflection of the population composition (where 80% of the population and 80% of students are white for example). The segregation ratio is, therefore, one. But there could still be considerable inequality in the system if the two groups were disproportionately represented in different universities (in the old and the post-1992 universities for example). What is also needed is a

measure of how evenly distributed the two groups are across all institutions. One formula (for the segregation index, see Gorard and Fitz 2000) is:  $S = \sum [|e_i/E - n_i/N|]/2$ . Here we are considering the pattern of distribution between the universities, rather than simply comparing the population of all universities with the total population. The value  $e_i$  is the number of ethnic minority students in University  $i$ ,  $E$  is the total number of ethnic minority students in all universities,  $n_i$  is the number of students in University  $i$ , and  $N$  is the total number of students in all universities. The  $| |$  symbols mean that we are interested only in the absolute difference between  $e_i/E$  and  $n_i/N$  (termed the residual), ignoring negative signs. The  $\sum$  symbol represents the sum of this difference for all cases.

Using the hypothetical values in Table 4, the segregation index is half the sum of the residuals (ignoring their signs).  $100/600$  minus  $20/300$  equals  $0.1$  and so on. The total of these residuals is  $0.33$  (one third) and half of that is around  $0.17$ . This is a measure of how segregated this imaginary university system is. Strictly speaking it is the 'exchange proportion', which is the proportion of the ethnic minority students who would have to be exchanged with white students in order to achieve a perfectly balanced distribution in all universities. In this example, half of all students are classified as ethnic minority. If these were evenly distributed then University 1 would have 50 ethnic minority students (not 20), University 2 would have 100 (not 80), and University 3 would have 150 (not 200). If the 50 'extra' ethnic minority students in University 3 were exchanged with 30 white students from University 1 and 20 from University 2, then there would be an even distribution of both white and ethnic minority students. Since  $50/300$  is equal to  $0.17$ , which is our calculated value for the segregation index, this tells us that the overall system is 17% segregated (or that 17% of ethnic minority students would have to be moved to eliminate all segregation). As with the segregation ratio, this approach can be used with other categories (male/female, pass/fail etc.), other organisational units (hospitals or occupations instead of universities for example), and any number of cases. It is a general measure of unevenness, which can be used reasonably safely for comparisons of inequality across time, place and other categories. It is, unlike most other approaches, strongly composition invariant (Gorard and Taylor 2002).

**Table 4 - Worked example of segregation index**

	Ethnic minority students	Total students	Residual
University 1	20	100	$100/600 - 20/300$
University 2	80	200	$200/600 - 80/300$
University 3	200	300	$200/300 - 300/600$
Total	300	600	0.33

As can be seen both of the indices described so far are based on a comparison between the proportion of one group and the proportion of the total population in each unit of analysis. These are seen to be the closest to what we mean when we talk about segregation or inequality. However, other indices have been proposed and some of these are in more common use (such as the dissimilarity index which compares the proportion of one group with the proportion of its inverse group). Since the 1930s many social scientists have been involved in 'index wars', fighting over the relative merits of each approach. Even measuring the strength of association in a simple two-by-two table gives rise to controversies that span generations, and still fascinate sociologists of science today (MacKenzie 1999). Pioneers in statistics, such as Pearson and Yule, could not agree how to perform this (apparently) simplest of calculations. The importance of this is that the precise nature of our findings is often dependent on our selection of an appropriate technique, which requires a deeper consideration of what precisely we are trying to measure. Unfortunately, all too often, this lack of agreement on how to present numeric findings is used to bamboozle readers, for work involving numbers often has considerable rhetorical power – a power that we submit to unless we are prepared to get involved in critique ourselves.

### **The rhetorical power of numbers**

According to Gigerenzer (2002) the Mexican government once increased a four-lane motorway to six narrower ones, and announced this as a 50% increase in capacity. When this led to more accidents, they reduced the lanes back to four, which was a 33% reduction in capacity. Thus, they claimed they were still 17% up on the deal! This is similar to the story of a company that loses business (e.g. from a base of 100 to 50), and then regains some of it (to 75). Their shareholders are told that the first period

led to a 50% decrease but that the second to a 50% increase. So, the company is back where it started. Of course, both of these examples are silly, but they show the potential for misleading others with numbers. There are two lessons. We should make our use of numbers as transparent as possible, and we should all be prepared to make the effort to understand and, if necessary, critique the work of others involving numbers.

Figures for the USA in 1999 suggest that around 25% of adults smoke, and that around 1 in 10 of these develops lung cancer compared to around 1 in 200 of non-smokers. What is the strength of the relationship between smoking and lung cancer? If we were to sample 800 adults we might get 200 smokers (25%) of whom 10% get lung cancer (Table 5)

**Table 5 – relationship between smoking and lung cancer 1**

	cancer	not	total
smoker	20	180	200
non-smoker	3	597	600
total	23	777	800

If, on the other hand, we sample 400 smokers and 400 non-smokers, we might get the figures in Table 6.

**Table 6 – relationship between smoking and lung cancer 2**

	cancer	not	total
smoker	40	360	400
non-smoker	2	398	400
total	42	758	800

Finally, if we sample 400 people with lung cancer and 400 without, then we might get the figures in Table 7.

**Table 7 – relationship between smoking and lung cancer 3**

	cancer	not	total
smoker	348	93	441
non-smoker	52	307	359
total	400	400	800

Note that each of these tables, quite properly, has the same row and column headings. The figures within them are based on the same frequencies. But each gives an entirely different impression. Most notably, the first two tables appear to suggest that smoking is considerably more benign than the third. The relationship (diagonal) appears stronger after the event (working back from cancer) than before (working forward from smoking). Which version is used might depend on the prejudice of the researcher, or their funder.

A study of 280,000 women in Sweden assessed the impact of a screening programme for breast cancer. Over ten years, there were 4 deaths per 1,000 among the over 40s without screening, and 3 deaths per 1,000 with screening. There are several different ways the gain from screening could be expressed. The absolute risk reduction is 1 in 1,000 or 0.1%. The number needed to treat (to save one life) is 1,000. The increase in average life expectancy is 12 days. And the relative risk reduction is 25%. It is the last that is routinely used by advocates and those standing to gain from screening programmes, perhaps because it sounds like a saving of 25 lives per 100. All versions are correct – but the relative risk is more likely to get funding and headlines. Information leaflets about the screening procedure mostly do not discuss false positives or other costs. Some even give the illusion that screening can reduce the incidence of the cancer. But to achieve even the level of success above requires about 9 in 10 false positives, and the distress and unnecessary operations that these entail. To these we must add the danger of cancer from the screening itself, and the financial cost of the programme (and therefore the lost opportunity to spend this money on reducing the risk in some other way). So viewed dispassionately, and with alternative ways of looking at the same data, a 1/1000 risk reduction may not seem worth it for this group. A similar issue of risk reduction, and the opportunity cost that it entailed, arose in the UK in 2002 concerning security checks on adults working with children. We cannot do the

risk reduction calculation for this yet since no figures have been published for the number of problem cases uncovered compared to the number of cases examined.

We routinely face scares about health, food, pesticides, the environment, and education of course, with the figures presented by their peddlers in the most rhetorically convincing way possible. Media reports of epidemiological studies tend to use the rhetorical power of numbers to make research reports appear more alarming or more flattering. We are unlikely to be able to change this directly. But our ability to see beyond the presentation, and to consider other equally valid presentations *is* under our control. Improving the ability of the consumer is our best defence, and will harm only those for whom the ideology is more important than the evidence (or who stand to benefit in some way from confusion). This improvement will be *most* marked in the apparently simple kinds of problems discussed so far. The lessons could perhaps be summarised as: be aware of the potential weakness of the measurements themselves, demand a base rate or comparison group for any claim, and be prepared to re-work the figures yourself in different ways to lessen their rhetorical power.

### **What is a statistical test?**

The next section of the paper moves beyond issues of measuring, counting and proportions to consider that bugbear of many ‘quantitative methods’ courses – statistical tests. We try to show that the logic of these tests is, in reality, quite simple and within the capacity of all novice researchers to grasp. There is often actually little need for such tests, but in current practice and due to general ignorance of their limitations they retain considerable rhetorical power.

Suppose that one of the background questions in a survey using a random sample of 100 adult residents in one city asked for the sex of the respondent. The results might be presented as in Table 8.

### **Table 8 - Frequency by sex in our achieved sample**

	Number
Male	41
Female	59
Total	100

Suppose that one of the substantive questions in the same survey asked the respondents whether they had received any work-based training in the past two years. The results might be presented as in Table 9.

**Table 9 - Frequency of training in our achieved sample**

	Number
Training past two years	53
No training past two years	47
Total	100

We know, therefore, that our achieved sample contained more women than men, and that slightly more than half reported receiving training in the past two years. Both of these might be important findings given a good-quality sample of a clearly defined population. In many cases, however, our chief concern as social scientists is to go beyond these simple patterns and answer questions such as 'Are men or women more likely to report receiving training?'. In this case we need to consider the two variables simultaneously, and we can present our summary as a cross-tabulation using different rows for one variable, and different columns for the other (Table 10).

**Table 10 - Cross-tabulation of sex by receiving training**

	Training	No training	Total
Male	24	17	41
Female	29	30	59
Total	53	47	100

Note that the 'marginal' totals are the same as in the simpler tables above. There are still 100 cases of whom: 41 are male; 53 received training; and so on. The table also now shows that slightly more than half of the men received training (24/41 or around 59%), while fewer than half of the women did (29/59 or 49%). For our sample, therefore, we can draw safe conclusions about the relative prevalence of reported training in the two sex categories. The men in our sample are more likely to have received training. However, our main motive for using probability sampling was that we could then generalise from our sample to the larger population for the study. If the population in this example is *all* residents of the city from which the sample was taken, can we generalise from our sample finding about the relationship between sex and receiving training? Put another way, is it likely that men in the *city* (and not just in the sample) were more likely to receive training than women?

In order to answer the question for the population (and not just for the people we asked by selecting them at random) it is very useful to imagine that the answer is 'no' and start our consideration from there (this is known as our 'null hypothesis'). If the answer were actually no, and men and women were equally likely to report training, then what would we expect the figures in Table 10 to look like? The number of each sex remains as defined in Table 8, and the number of people receiving training remains as defined in Table 9. In other words our table of what we would expect to find starts with the partially completed Table 11.

**Table 11 - The marginal totals of sex by training**

	Training	No training	Total
Male			41
Female			59
Total	53	47	100

From this outline we can calculate exactly what we expect the numbers in the blank cells to be. We know that 41% of cases are male, and that 53% of cases received training. We would therefore expect 41% of 53% of the overall total to be both male and have received training. This works out at around 22%, or 22 cases as shown in Table 12.

**Table 12 - The expected value for males receiving training**

	Training	No training	Total
Male	22		41
Female			59
Total	53	47	100

We can do the same calculation for each cell of the table. For example, as 59% of the cases are female and 53% of the cases received training, we would expect 59% of 53% of the overall total to be females and have received training. This works out at around 31%, or 31 cases. But then we already knew that this must be so, since 53 people in our survey received training, of whom we expected 22 to be male, so by definition we expected the other 31 to be female. Similarly, 41 cases are male and we expected 22 of these to have received training, so we expected that the other 19 did not. In technical terms, this means that the ‘degrees of freedom’ in the table are one. We can now complete the table (Table 13). Note that in practice all of these calculations would be generated automatically by a computer.

**Table 13 - The expected values for sex by receiving training**

	Training	No training	Total
Male	22	19	41
Female	31	28	59
Total	53	47	100

To recap, we obtained the figures in Table 10 from our survey (our 'observed' figures) and wanted to know whether the apparent difference in training rates for men and women was also likely to be true of the city as a whole. To work this out, we calculated how many men and women we expect to have received training assuming that there was actually no difference, and obtained the figures in Table 13 (our 'expected' figures'). For convenience both sets of figures are repeated in Table 14 with the observed figures in each cell followed by the expected figure in brackets.

**Table 14 - Observed and expected values for sex by receiving training**

	Training	No training	Total
Male	24 (22)	17 (19)	41
Female	29 (31)	30 (28)	59
Total	53	47	100

If there were no difference in the city as a whole between the rates of receiving training for men and women then we would expect 22 of 41 males to have received training, but we actually found 24 of them. In each cell of the table there is a discrepancy of two cases between the observed and expected figures. Is this convincing evidence that men are more likely than women in this city to report receiving training? Hardly. In selecting a sample of 100 cases at random it would be easy for us to have inadvertently introduced a bias equivalent to those two cases. We should therefore conclude that we have no evidence of differential training rates for men and women in this city.

The argument traced in Tables 10 to 14 contains just about everything that you need to know about the logic of significance-testing in statistical analysis. It is a form of logic that all social science researchers should be able to follow. If you can follow the logic, and are happy with the conclusion, then you have completed a statistical analysis. There is, therefore, no reason why anyone should not read and understand statistical evidence of this sort. Everything about traditional statistics is built on this rather simple foundation, and yet nothing in statistics is more complicated than this. Much of what appears in traditional texts is simply the introduction of a technical shorthand for the concepts and techniques used in this introductory argument (see Gorard 2003). The key thing a statistical test does for us is to estimate how unlikely it is that what we observed would happen, assuming, for the sake of calculation, that the null hypothesis is true. The more unlikely our observation is then the less likely it is that the null hypothesis is true, and therefore the more likely that we need to look for an alternative explanation. The result is usually expressed as a probability (the probability of the null hypothesis being true is around 30% in this example), but as in the earlier examples we believe that the underlying logic is easier for beginners to see when expressed as frequencies as they are here.

## **Do we need statistical tests?**

This form of statistical testing has many historical roots, although many of the tests in common use today, such as those attributable to Fisher, were derived from agricultural studies (Porter 1986). They were developed for one-off use, in situations where the measurement error was negligible, in order to allow researchers to estimate the probability that two random samples drawn from the same population would have divergent measurements (men and women in the example above). In a roundabout way, this probability was then used to help decide whether the two samples actually come from two different populations. For example, vegetative reproduction could be used to create two colonies of what is effectively the same plant. One colony could be given an agricultural treatment, and the results (in terms of survival rates for example) compared between the two colonies. Statistics would help us estimate the probability that a sample of scores from each colony would diverge by the amount we actually observe, assuming that the treatment given to one colony was ineffective. If this probability is very small, therefore, we might conclude that the treatment appeared to have an effect. As we have seen, that, in a nutshell, is what significance tests are, and what they can do for us.

In light of current practice, it is important to emphasise what significance tests are not, and cannot do for us. Most simply, they cannot make a decision for us. The probabilities they generate are only estimates, and they are, after all, only probabilities. Standard limits for retaining or rejecting our null hypothesis of no difference between the two colonies, such as 5%, have no mathematical or empirical relevance. They are only arbitrary. A host of factors might affect our confidence in the probability estimate, or the dangers of deciding wrongly in one way or another. Therefore there can, and should, be no universal standard. Each case must be judged on its merits. However, it is also often the case that we do not need a significance test to help us decide this (as in the training example above). In the agricultural example, if all of the treated plants died and all of the others survived (or vice versa) then we do not need a significance test to tell us that the probability is very low (and precisely how low depends on the number of plants involved) that the treatment had no effect. If there were 1,000 plants in the sample for each colony, and one survived in the treated group, and one died in the other group, then again a significance test would be superfluous (and so on). All that the test is doing is formalising the estimates of relative

probability that we make anyway in everyday situations. They are really only needed when the decision is not clear-cut (for example where 600/1000 survived in the treated group but only 550/1000 survived in the control), and since they do not make the decision for us, they are of limited practical use even then. Above all, significance tests give no idea about the real importance of the difference we observe. A large enough sample can be used to reject almost any null hypothesis on the basis of a very small 'effect' (see below).

It is also important to re-emphasise that the probabilities generated by significance tests are based on random samples. If the researcher does not use a random sample then inferential statistics are of little use since the probabilities become meaningless. Researchers using significance tests with convenience, quota or snowball samples, for example, are making a key category mistake. Similarly, researchers using significance tests on populations (from official statistics perhaps) are generating meaningless probabilities. It is possible that a trawl of education journals would reveal very few, technically, correct uses of significance tests. Added to this is the problem that social scientists are not generally dealing with variables, such as plant survival rates, with minimal measurement error. In fact, many studies are based on latent variables, such as attitudes, of whose existence we cannot even be certain, let alone how to measure them (see section above on measuring length). Added to this are the problems of non-response and participant dropout in social investigations, that also do not occur in the same way in agricultural applications. All of this means that the variation in observed measurements due to the chance factor of sampling (which is all that significance tests take into account) is generally far less than the potential variance due to other factors (see section above on ill-conditioning). The probability from a test contains the unwritten proviso - assuming that the sample is random with full response, no dropout, and no measurement error. The number of social science studies meeting this proviso is very small indeed. To this must be added the caution that probabilities interact, and that most analyses in the IT age are no longer one-off. Most analysts start each probability calculation as though nothing prior is known, whereas it may be more realistic and cumulative to build the results of previous work into new calculations (Roberts 2002).

Significance tests have a specific valuable role to play in a limited range of research situations. Therefore, while it is important for novice social scientists to be taught about the use of significance tests,

it is equally important that they are taught about the limitations as well (and alerted to possible alternatives, such as confidence intervals, effect sizes, and graphical approaches). But even these alternative statistics cannot be used *post hoc* to overcome design problems or deficiencies in datasets. If all of the treated plants in our example were placed on the lighter side of the greenhouse, with the control group on the other side, then the most sophisticated statistical analysis in the world could not overcome that bias. It is worth stating this because of the current push for more complex methods of probability-based analysis when a more fruitful avenue for long-term progress would be the generation of better data, open to inspection through simpler and more transparent methods of accounting. Without adequate empirical information 'to attempt to calculate chances is to convert mere ignorance into dangerous error by clothing it in the garb of knowledge' (Mills 1843, in Porter 1986, p.82-83). Null hypothesis significance tests (NHSTs) may therefore be a hindrance to scientific progress (Harlow et al. 1997).

Statistics is not, and should not be, reduced to a set of mechanical dichotomous decisions around a 'sacred' value such as 5%. Suggested alternatives to reporting NHSTs have been the use of more non-sampled work, effect sizes (Fitz-Gibbon 1985), meta-analyses, parameter estimation (Howard et al. 2000), or standard confidence intervals for results instead, or the use of more subjective judgements of the worth of results. In the US there has been a debate over whether the reporting of significance tests should be banned from journals to encourage the growth of these alternatives (Thompson 2002). Both the American Psychological Society and the American Psychological Association have recommended reporting effect sizes and confidence intervals, and advocated the greater use of graphical approaches to examine data. Whereas a significance test is used to reject a null hypothesis, an effect size is an estimate of the scale of divergence from the null hypothesis. The larger the effect size, the more important the result. A confidence interval may be defined by a high and low limit between which we can be 95% confident (for example) that the 'true' value of our population estimate lies. The smaller the confidence interval the better quality the estimate is (de Vaus 2002).

Of course, several of the proposed replacements, including confidence intervals, are based on the same sort of probability calculations as significance tests. Therefore, they are still largely inappropriate for use with populations and non-random samples, and like significance tests they do nothing to overcome

design bias or non-response. Most of the alternatives require considerable subjective judgement in interpretation anyway. For example, a standard effect size from a simple experiment might be calculated as the difference between the mean scores of the treatment and control groups proportional to the variance (or the standard deviation) for that score among the population. This sounds fine in principle, but in practice we will not know the population variance. If we had the population figures then we would not need to be doing this kind of calculation anyway! We *could* estimate the population variance in some way from the figures for the two groups, but this introduces a new source of error, and the cost may therefore over-ride the benefit on several occasions. There is at present no clear agreement, other than the need for the continued use of intelligent judgement.

Recent UK initiatives, perhaps most prominently the new funding arrangements for ESRC PhD students, have been designed to encourage a wider awareness of statistical techniques among social scientists. While these moves are welcome, the lack of agreement about the alternatives, the absence of textbooks dealing with them (Curtis and Araki 2002), and their need for greater skill and judgement means there is a consequent danger of simply re-visiting all of the debates about statistics that have taken place in other disciplines since at least 1994 (Howard et al. 2000). Although there are suggestions to replace probability values with standard errors or confidence intervals (e.g. Altman et al. 2000), many of the same problems would continue to apply. It is not clear why we should use standard errors anyway. They are not used in business reports, or examination grades, for example, where they might be just as appropriate. In real-life the best estimate is our current score for any measurement (while we should treat all such scores with caution).

## **Conclusion**

Part of what this paper tries to do is show that standard approaches to significance testing, currently the cornerstone of many 'quantitative' methods courses, should no longer have automatic pride of place. There is a pressing need for more general awareness of the relatively simple role of numbers in those common social scientific situations for which probabilities are not relevant. The importance of this ongoing debate about tests is that it suggests that we need to move away from a formulaic approach to

research. However, we need to replace empty formulae for reporting results, not with an 'anything goes' philosophy, but with almost anything goes as long as it can be described, justified and replicated. Above all, we need to remember that statistical analysis is not our final objective, but the starting point of the more interesting social science that follows. A 'significant' result is worth very little in real terms, and certainly does not enable us to generalise safely beyond a poor sample. The key issue in research is not about significance but about the quality of the research design.

More generally, much of what we know about the social world is uncertain. Our knowledge, such as it is, is commonly expressed in terms of probabilities. The same situation applies to health research, engineering, judicial proceedings and all the other fields represented in the examples used above. All professionals have their own craft knowledge, which is at least partially derived from research evidence. However, in a number of practical situations, more explicit use of research findings has been shown to lead to more beneficial outcomes than relying solely on professional experience. The practical problem is that these research findings are generally expressed in terms of risk reduction and uncertainty. If the probabilities are misunderstood by the professionals, as they have been in several of the examples in this paper, then it is difficult for the professionals to make their own judgements in practice, and the use of research findings could, in this situation, have far from beneficial outcomes. In building the capacity to generate and use research evidence relevant to teaching and learning in the UK, it might not be an exaggeration to say that what we need above all else is a more general willingness and ability among interested parties to think about the nature of uncertainty.

## **References**

- Altman, D., Machin, D., Bryant, T. and Gardiner, M. (2000) *Statistics with confidence*, London: BMJ Books
- Brown, A. and Dowling, P. (1998) *Doing research/Reading research: a mode of interrogation for education*, London: Falmer
- Curtis, D. and Araki, C. (2002) *Effect size statistics: an analysis of statistics textbooks*, presentation at AERA, New Orleans April 2002
- Dawes, R. (2001) *Everyday irrationality*, Oxford: Westview Press
- de Vaus, D. (2001) *Research design in social science*, London: Sage
- Fitz-Gibbon, C. (1985) The implications of meta-analysis for educational research, *British Educational Research Journal*, 11, 1, 45-49
- Fleiss, J. (1973) *Statistical methods for rates and proportions*, New York: John Wiley and Sons
- Gigerenzer, G. (2002) *Reckoning with risk*, London: Penguin
- Gorard, S. (2001) *Quantitative methods in educational research: the role of numbers made easy*, London: Continuum
- Gorard, S. (2002) Combining methods, *Research Papers in Education*, 17 (forthcoming)
- Gorard, S. (2003) *The role of numbers in social science research: quantitative methods made easy*, London: Continuum
- Gorard, S. and Fitz, J. (2000) Investigating the determinants of segregation between schools, *Research Papers in Education*, 15, 2, 115-132
- Gorard, S. and Taylor, C. (2002) What is segregation? A comparison of measures in terms of strong and weak compositional invariance, *Sociology*, 36, 5 (forthcoming)
- Gorard, S., Salisbury, J. and Rees, G. (1999) Reappraising the apparent underachievement of boys at school, *Gender and Education*, 11, 4, 441-454
- Harlow, L., Mulaik, S. and Steiger, J. (1997) *What if there were no significance tests?*, Marwah, NJ: Lawrence Erlbaum
- Heath, A. (2000) The political arithmetic tradition in the sociology of education, *Oxford Review of Education*, 26, 3&4, 313-331

- Howard, G., Maxwell, S. and Fleming, K. (2000) The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis, *Psychological Methods*, 5, 3, 315-332
- MacKenzie, D. (1999a) The science wars and the past's quiet voices, *Social Studies of Science*, 29, 2, 199-213
- Plewis, I. (1997) Presenting educational data: cause for concern, *Research Intelligence*, 61, p.9-10
- Porter, T. (1986) *The rise of statistical thinking*, Princeton: Princeton University Press
- Prandy, K. (2002) Measuring quantities: the qualitative foundation of quantity, *Building Research Capacity*, 2, 3-4
- Roberts, K. (2002) Belief and subjectivity in research: an introduction to Bayesian theory, *Building Research Capacity*, 3, 5-6
- Schagen, I. (2002) Well, what do you know?, *Education Journal*, 63, pp.27-28
- Siegel, S. (1956) *Nonparametric Statistics*, Tokyo: McGraw-Hill
- Thompson, B. (2002) What future quantitative social science could look like: confidence intervals for effect sizes, *Educational Researcher*, 31, 3, 25-32
- Wrigley, J. (1976) Pitfalls in educational research, *Research Intelligence*, Vol. 2, No. 2, p.2-4

### **Recommended further reading**

- Berka, K (1983) *Measurement: its concepts, theories and problems*, London: Reidel
- Booth, W., Colomb, G. and Williams, J. (1995) *The craft of research*, Chicago: University of Chicago Press
- Brignell, J. (2000) *Sorry, wrong number! The abuse of measurement*, European Science and Environment Forum
- Clegg, F. (1992) *Simple Statistics: a course book for the social sciences*, Cambridge: Cambridge University Press
- Dawes, R. (2001) *Everyday irrationality*, Oxford: Westview Press
- Gephart, R. (1988) *Ethnostatistics: Qualitative foundations for quantitative research*, London: Sage

Gigerenzer, G. (2002) *Reckoning with risk*, London: Penguin

Gorard, S. (2003) *The role of numbers in social science research: quantitative methods made easy*,  
London: Continuum

Huck, S. and Sandler, H., (1979) *Rival hypotheses: Alternative interpretations of data based  
conclusions*, New York: Harper and Row

Huff, D. (1991) *How to lie with statistics*, Harmondsworth: Penguin

Reichmann, W. (1961) *Use and abuse of statistics*, Harmondsworth: Penguin

Thouless, R. (1974) *Straight and crooked thinking*, London: Pan



ISBN 1 872330 99 1