A Simple Guide to Voodoo Statistics

Ian Schagen
Chief Research Analyst
New Zealand Ministry of Education

"There are lies, damn lies - and statistics." (Mark Twain, attributed to Disraeli)

In this discursive, but hopefully brief, paper, I intend to explore a number of issues:

- How do we ensure that policy and practice in education is well-served by statistical information?
- How do we resolve the tension between simplicity of analysis and providing misleading statistics?
- What should be the place of statistical literacy in 21st century education?

This paper is prompted by over 20 years working as a statistician within the field of educational research. If it is true that we learn by our mistakes, then I can claim to have learned a great deal in that time. I have also encountered a great deal of what I have termed 'voodoo statistics' – numerical results, measures and indicators that are not fit for purpose and conceal rather more than they reveal. In the next section I will try to analyse some of these examples of voodoo statistics – not in a pejorative way, but to see what we can learn to ensure that things get better.

Examples of Voodoo Statistics

1. School performance tables

Inevitably the best recent example within the English educational system is the use of raw examination and test outcomes to quantify school performance. This was on a par with judging shops purely on their turnover, irrespective of size, location or other factors such as what they were selling. Despite this, the power of numbers was such that it was widely assumed (and still is in some quarters) that a school in a well-heeled suburb, with a high performing intake and with 52% achieving 5+ A to C grades at GCSE, was somehow 'doing better' than another school with multiple hits from deprivation, low performing intake etc., which achieved 15% 5+ A to C grades. In fact, the quoted figures told us a lot about the school's intake and environment and next to nothing about its actual performance as a school.

How did this happen? What persuaded intelligent people that these voodoo statistics actually had the meaning they were credited with? I can only suggest one or two possible explanations:

- There was a real drive for school accountability, which had been absent up to that point, and raw attainment measures were seized on as giving some kind of measure which was relevant, for want of anything better.
- Simplicity is strongly valued for its own sake, with an underlying assumption that a plain and straightforward measure will tell us all we need to know. In some cases this is true, but in many it is not. A previous Secretary of State for Education (Kenneth Clarke) famously remarked that he liked his data raw, not cooked.
- Perhaps there was an assumption that anything more complex would not be understood by the average parent or teacher, despite the fact that most people accept information displayed as, for example, a Retail Price Index or a weather forecast chart – probably only because they have confidence in the expertise of those who produce them.

2. "<Insert Type here> schools are improving faster"

This follows on from the above: take a voodoo statistic of school performance and measure its year-on-year change for a particular group of schools compared with the national average. Here are only a few recent examples: "Exam results in academy schools are improving faster than the national average"[1]; "Healthy schools 'improve faster'"[2]; "GOL helps London schools improve faster than the national average"[3]. Without wishing to rain on anyone's parade, there are a number of statistical problems with this kind of stuff, quite apart from the problem of measuring school performance on raw outcomes outlined above.

To illustrate this issue, and how things can go horribly wrong, I have created a simple statistical model. Suppose we have two populations, 'national' and 'Type X', and both have normally distributed values on some outcome. In Year 1 the national group has mean 5.0 and standard deviation 1.0, while the Type X group has mean 3.5 and the same standard deviation. Each year the former group's mean increases by 0.25, while the latter group's mean increases by 0.2, with standard deviations held constant. Suppose further that the 'threshold' of interest is a score of 5.0 or more, and the indicator is the percentage in each group achieving this.

Figure 1 is the headline indication for change over time, as espoused by the headline makers. It shows the percentage passing the threshold for both groups, as a percentage of the original percentage in Year 1. This is certainly impressive – compared with the national group, the Type X group is making staggering progress: nearly 50% in Year 2, and approaching 300% improvement in Year 5!

"Inevitably the best recent example [of voodoo statistics] within the English educational system is the use of raw examination and test outcomes to quantify school performance."

Figure 2 below shows the more mundane reality – that the mean scores for both groups are rising steadily, but the Type X group is rising at a slightly slower rate.

Why this discrepancy? Figure 3 shows one of the intermediate stages between the above two diagrams – it shows the actual percentages achieving the threshold for the two groups each year. Both are rising, and if anything the national group seems to be rising faster. However, when we present these percentages as percentages of the values in Year 1, we get Figure 1, which is highly sensitive to the initial values. In fact, Figure 1 is extremely misleading compared with either of the other two plots.

Another reason why such statistics should be regarded with suspicion is the well-known statistical phenomenon of 'regression to the mean'. Put briefly, this states that if we select a sample of cases whose values are in general higher or lower than the population mean, and measure those values again after a period of time, then there will be a tendency for the new values to be

closer to the mean than the original measurements. To see how this works try the following simple experiment: Take a population of schools, children or whatever and measure them all on some criterion (% 5+ A*-C grades, mathematics test score, etc.).

- Select the bottom 10% of the population and apply an arbitrary intervention – for example, paint them pink.
- A year later re-test the same population. In general the pink-painted 10% will have improved since last time – some of them will no longer be in the bottom 10%, and some of those in the bottom 10% will not be pink.
- Interpret these results to demonstrate that your intervention ("Paint-'Em-Pink') has had a transforming effect on outcomes for the bottom 10%.

Unfortunately, the same experiment in reverse, applied to the top 10%, will show conclusively that it has a negative effect on high performers. This kind of effect is virtually inevitable when dealing with individuals who change over time in different directions – those at the bottom are more likely to go up, and those at the top more likely to go down. It is important to appreciate this before setting up evaluations of interventions on extreme cases which are guaranteed to show a positive effect.

3. Mind the Gap

'Narrowing the Gap' is one of the current concerns of many within government and policy. This is illustrated by articles from the Child Poverty Action Group[4]; headlines about "£1bn to close poverty results gap"[5]; and articles about the 'gender gap' (sic)[6]. In many ways this is a welcome development, focusing on apparent disparities in outcomes between various apparently vulnerable groups and their peers, and seeking to find ways to redress these. However, there are some important statistical caveats that must be borne in mind during all this activity, and some messages that may be worth taking on board from a careful analysis of the data.

Before carrying out any work on 'narrowing the gap', it is important for us to be clear on what exactly 'the gap' consists of. The expression calls to mind the concept of a distinct break in outcomes, with children in a particular vulnerable group all on one side of a great divide, and everyone else on the

Figure 1: Illustration of the % change in % passing the threshold

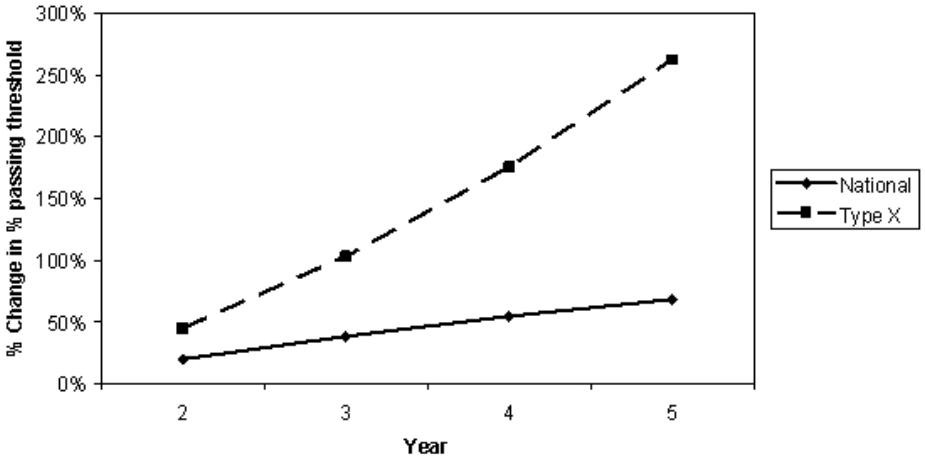


Figure 2: Changes in underlying mean over five years

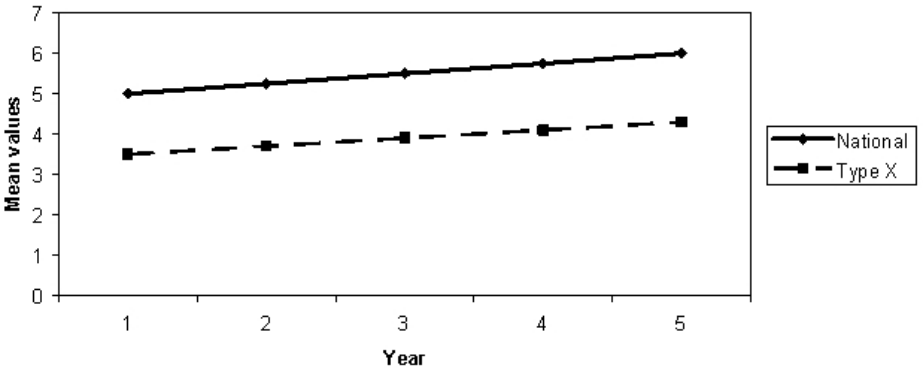


Figure 3: Percentages passing the threshold over time

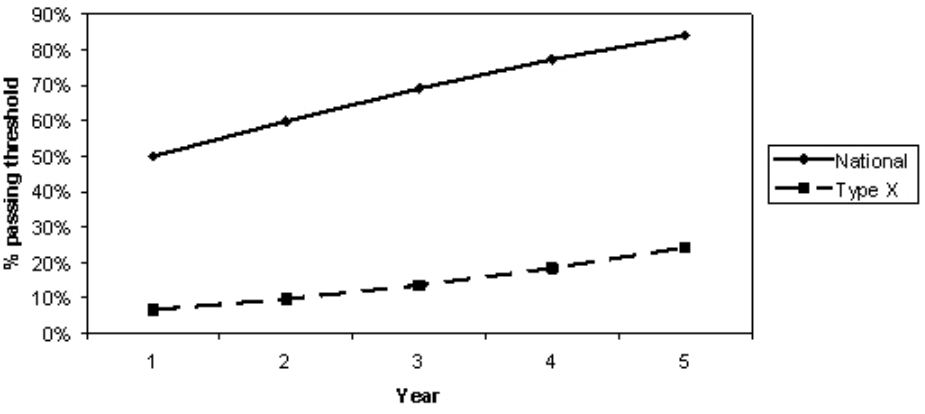
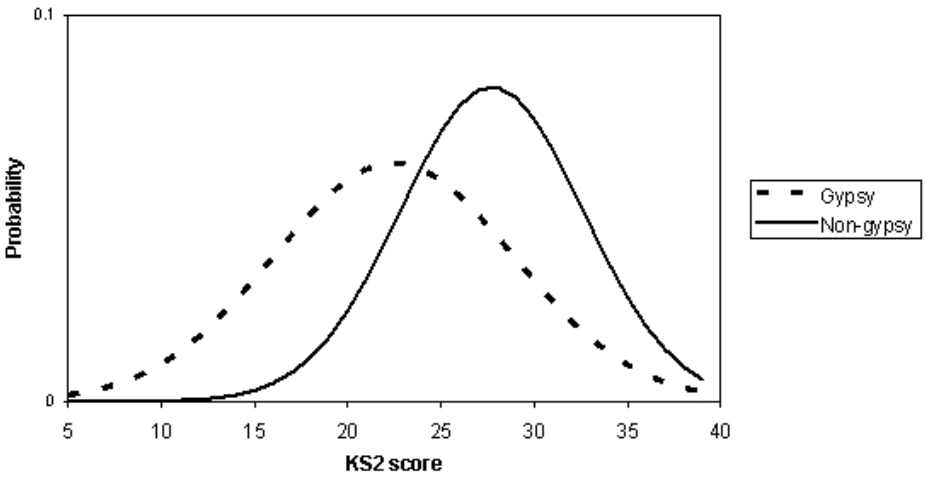


Figure 4: KS2 score distributions – Gypsy and non-Gypsy



other side. This is clearly not the case – within any group there is a wide dispersion in outcomes, and a great deal of overlap between that group and others. To make this discussion more concrete, let us consider a particular ethnic group, say gypsy children, and look at their performance at KS2 in 2006 compared with all others. Looking at the two populations, gypsy pupils and the rest, there is a 'gap' of 5.26 points based purely on raw KS2 scores, with the mean for gypsy pupils being this much below the mean for the rest. This is illustrated in Figure 4 below, which also shows the large amount of overlap between the two populations.

We are right to be concerned about the gap between gypsy pupils and the rest, but it is also important to realise that neither group is a homogeneous mass. Within each group there are quite wide disparities between the highest and lowest performers, and these may need to be explained by factors other than gypsy/non-gypsy.

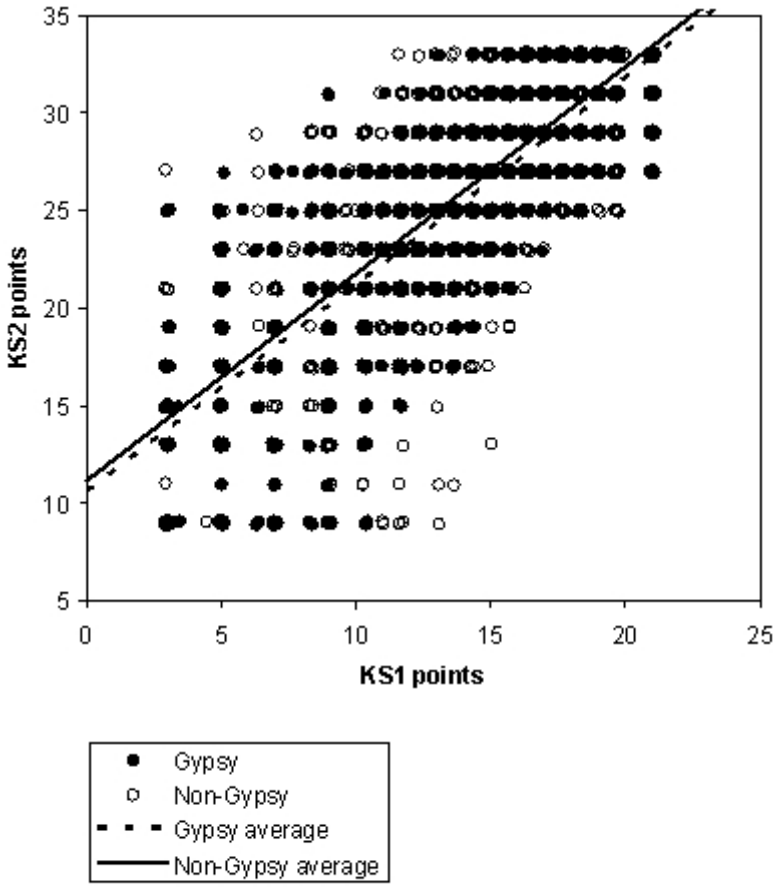
"If the attainment results for individual pupils are to a large extent unpredictable and do not fall into neat groups with a clear 'gap', then this is almost certainly the case for other Every Child Matters (ECM) outcomes."

The size of the 'gap' also depends on what else we take into account, as illustrated by Figure 5 below, which shows key stage 2 results plotted against prior attainment at key stage 1, for the two groups: minority ethnic gypsy children, and (a sample of) all others. The plot also shows the difference in average attainment for the two groups.

If we take account of pre-existing differences at KS1, the size of the 'gap' goes down to 0.54 points, about 10% of the original. If we take account of other factors, for instance eligibility for free school meals, then the size of the 'gap' is reduced further, to 0.40 points on average.

If the attainment results for individual pupils are to a large extent unpredictable and do not fall into neat groups with a clear 'gap', then this is almost certainly the case for other Every Child Matters (ECM) outcomes, such as health, safety, enjoyment, etc. This is not to deny that there are tendencies for certain groups to have, on average, lower outcomes than others, but this may partly be because there are prior characteristics for such groups that reinforce these tendencies. One of the questions we need to face is the extent to which vulnerable groups have lower than expected outcomes, allowing for such background characteristics. The challenge for those who want to make positive change is to understand the complex inter-relationships between factors and to target interventions in a sophisticated way, rather than assume that all members of a group have the same characteristics and need the same kind of treatment.

Figure 5: KS1 to KS2 progress – Gypsy pupils vs non-Gypsy



This is not to deny the importance of ensuring good outcomes for vulnerable groups, but analysis focusing solely on the 'gap' without looking in detail at all the other factors is likely to be misleading and unproductive.

How do we make things better?

It is fairly easy to find examples of the misuse of statistics, but in this paper I want to be positive and not just critical. I think there are some definite steps that everyone – statisticians, politicians, policy makers and practitioners – can do to stamp out voodoo statistics when they occur and ensure that we focus on analyses that are fit for purpose.

1. Make validity the chief yardstick rather than simplicity;
2. Do proper trials of new initiatives whenever possible;
3. Present statistical results clearly and accessibly;
4. Be objective and open, and earn trust;
5. Encourage statistical literacy.

There are certain things in life that are simple, like the law of universal gravitation, but their ramifications swiftly become quite complex. In educational and social research few things are very simple, and it is a mistake to use an indicator that is invalid just because it is simple. In practice, a range of indicators is usually available, each with its own strengths and weaknesses, and any measure we choose should be acknowledged to be imperfect and partial.

An example is the use of 'threshold' indicators, which convert a continuous range of values into a binary pass/fail measure. Sometimes we need such measures – we need to know if someone is fit to drive a car, or carry out brain surgery, or deserves a PhD. On many occasions, however, the division is arbitrary and the underlying continuous value is a more useful measure than the threshold indicator. At some stage it was decided that five GCSE passes at grades A* to C was an indicator of a 'good' outcome at age 16 – but it could have been 4 or 6, and

included grade B upwards only, giving quite different results. In the interests of simplicity we have collapsed a complex set of results for each individual into a single pass/fail indicator, but a lot of information has been lost in the process. Depending on the question being asked, different measures derived from the data can be used to provide clearer insights into the underlying, rather complex, situation.

Governments can give up launching new initiatives in the same way that bridge players can stop playing bridge, but there is a real need for them to become more restrained in what initiatives they launch and how they are evaluated. Too often they are based not on research but on politically-based assumptions about what 'must work', and put in place without a proper pilot or a clear plan for evaluation. Often they target a particular sector of the school population (e.g. under-performing schools or those in deprived areas) without leaving a comparable control group. As we have seen, regression to the mean can make such targeted initiatives appear successful even when they are having no impact.

Randomised control trials (RCTs) involve randomly allocating schools or individuals to 'treatment' groups (who receive the intervention) and 'control' groups (who do not), and are slowly becoming more common within social and educational research (see Mosteller and Boruch, 2002; Styles, 2006). In an ideal world, all educational initiatives would be piloted using an RCT to ensure they actually made a difference to desired outcomes before being implemented nationally. In stern political reality, the ideal may not be possible, but every effort should be made to carry out a pilot with the following properties:

- Treatment and control groups closely matched on relevant background characteristics even if not randomised;
- Robust and valid measures of the desired outcome collected initially and at a later stage post-intervention;
- Some kind of measure of the extent to which the intervention has been fully implemented in the treatment group, and whether there has been any 'leakage' into the control group;
- Careful and suitably sophisticated analysis of the data collected, looking not only for overall effects of the intervention but also differential effects for different groups;
- Clear and objective presentation of the results with no political 'spin'.

This kind of careful piloting may seem expensive and time-consuming to those who are impatient to see results from their pet initiative, but compared with implementing something ineffective nationally it is good value for time and money.

It is vitally important that statisticians can present the results of their analysis in a way that is valid and accessible to others, and this is not always easy. One of the problems is that academic journals seem, to a cynical mind, to have a policy on publishing papers: "If anyone can understand it, it's not clever enough". This leads statisticians to publish stuff that is quite opaque, full of equations, dense tables and even denser jargon. When we switch from academic papers to outputs for a wider audience, some of these bad habits can persist. Complex statistical models are often essential for understanding complex social data, but we need to find ways of presenting the key elements of our analysis in graphical, tabular or verbal forms that can speak directly to intelligent non-statisticians (see e.g. Schagen and Elliot, 2004).

> **"Too often [government initiatives] are based not on research but on politically-based assumptions about what 'must work', and put in place without a proper pilot or a clear plan for evaluation."**

The Twain/Disraeli quotation at the top of this paper illustrates the need to earn trust for statistical findings – and also that this is not a new problem. Every time some piece of voodoo statistics or politically-spun statistical analysis is shown to be invalid or selective with the data, trust in statistical analysis is reduced in the general public – hence the cry for 'uncooked' simplistic and invalid statistics. It is interesting that most of the commentators and media sources that use and rely on the Retail Price Index (RPI)[7] probably have little idea about how it is created and updated, yet trust it to be an accurate reflection of some underlying reality. If such a sophisticated economic indicator is felt to be reliable by its wider audience, why is this not the case in education? Perhaps there is a case for some careful research into this difficult area.

Statistical literacy[8] is one of the many 'literacies' advocated for living in the modern world, but I believe it is a key ability (see Schagen, 2006). It is nothing to do with being able to do mathematics, or even arithmetic, but about understanding the key concepts underpinning statistics – things like populations, samples, randomness, bias, uncertainty, estimates and errors. Once people begin to understand such ideas, they can ask critical questions of any statistics presented to them. Hopefully they will then be able to unmask the voodoo statistics and focus on those based on high-quality analysis, well-presented by those they trust without any spin.

The author was previously Head of Statistics at NFER, and in April 2008 became Chief Research Analyst at the New Zealand Ministry of Education. The views expressed in this article are his own, and do not represent the opinions of either organisation.

Footnotes

[1] See http://education.guardian.co.uk/newschools/story/0,,2131193,00.html
[2] See http://news.bbc.co.uk/2/hi/uk_news/education/4332967.stm
[3] See http://www.go-london.gov.uk/GOLNewsletter/october04/dfes_regional_visit.asp
[4] http://www.cpag.org.uk/info/Povertyarticles/Poverty123/gap.htm
[5] http://news.bbc.co.uk/1/hi/education/6662667.stm
[6] http://www.timesonline.co.uk/tol/life_and_style/education/article2034112.ece
[7] http://www.statistics.gov.uk/cci/nugget.asp?id=21
[8] See, for example, the International Statistical Literacy Project at http://course1.winona.edu/cblumberg/islphome.htm

References

Mosteller, F. and Boruch, R. (2002). *Evidence Matters; Randomized Trials in Education Research*. Washington DC: Brookings Institute.

Schagen, I. and Elliot, K. (Eds) (2004). *But What Does It Mean? The Use of Effect Sizes in Educational Research*. Slough: NFER.

Schagen, I. (2006). 'Statistical literacy is the essential skill for educational managers', *Education Journal*, 98, 21.

Styles, B. (2006). 'Educational research v. scientific research', *Research Intelligence*, 95, 7-9.