Jack Marwood's Festival of Education presentation, Friday 19th June 2015

I've called this '*What If Everything You Know About Education Data is Wrong?*' It is based on some writing which I have done for David Didau's book, *What If Everything You Know About Education is Wrong?* David's book asks you as a reader - and I'm going to ask you today - to consider the question, *"What if you are wrong?"* What if what you *think* you know about education data is wrong?

As I say in the book, "We live in an era of Big Data. If you are relatively new to teaching you may be surprised how little we used to know about schools. Until 1992, those not actually working in a given English secondary school had no way of finding out basic information about measures of achievement such as, say, the average exam results of sixteen year olds in a particular secondary school. Likewise, no one outside of an English primary school had access to any published data before 1996. There was simply no data publicly available to anyone outside of a school.

This changed through a combination of politics and computing power. The era of the personal computer has enabled huge amounts of data to be collected, disseminated and dissected. Governments have taken advantage of this, requiring schools to create and collect large amounts of information of considerably variable quality. Schools, particularly at secondary level, often employ specific data managers. Elsewhere this role is fulfilled by teachers or their managers. Modern schools are awash with a veritable ocean of data.

The thinking behind a great deal of education data is, however, flawed on multiple levels and is – in the main – completely wrong."

There are a number of things which I would like you to consider.

You might be a teacher, you might be somebody who has trained as teacher, you might currently be teaching, you might have taught. But if you trained as a teacher, I'd like to consider what you know about data. And what you know about education data.

So, for my relatively short time with you, I want you to consider some questions. These are the questions I'd like you to think hard about:

How much education data is actually Cargo Cult Data?

What should teachers know about psychometrics?

What should you know about education data?

What should we be doing?

By the end of the session I'd like you to have thought about what you know about numbers and counting, and the limits and possibilities of quantitative and qualitative data; I'd like you to think about temperature and how you can describe changes in temperature.

I'd like you to think about the difference between sampling data and population data; what a subsection of the population might do or say and how that might extrapolate to the whole of the population. The thought I'd like to put inside your mind is to consider the Polls prior to the recent general Election. I'd also like you to consider what you know about testing which is done under ideal

conditions, such as the testing which is done to find out how fuel efficient a particular car might be, and how that is done under optimum conditions by the manufacturer but what actually happens in real life – what happens with your car, for example – might be different.

So, how numbers work, how samples work and how real life isn't quite what you get under less than ideal conditions. All pollsters imply certainty, all manufacturers tell you that your car will achieve sixty miles to the gallon, but your experience is likely to be very different.

But primarily, by the end of the session I'd like you to think about what you know and therefore what you don't know and therefore how you might be wrong.

So. David asked me to consider data and it's something which I've been thinking and writing about for a while now. I'll give you a bit of my background. I'm a primary school teacher; I've been teaching for eleven years and I trained twelve years ago in 2003. In the training which I had, which was a PGCE which I came to after having worked in the corporate sector for a while, there was very little mention of data; it didn't feature highly, and that is largely because, even thirty years ago, data was a minimal part of education. It has since muscled onto centre stage. If you consider the different ways in which people get into education these days, however, if you consider routes such as School Centred ITT, Teach First, PGCE or B EDs, how much data is actually covered in that and therefore how much those who teach actually know about data?

My experience is that most people don't actually know much about numbers. Most people don't know much about data. People – particularly in Primary – haven't really been educated much beyond GCSE maths. It's very rare that you find anyone who has A Level maths; you do, but it isn't common. Most people seem to be using numbers intuitively, and that's a core focus of David's book: "What is our intuition and might it be wrong? What are our cognitive biases, what are the anchoring effects?"

Many teachers didn't really get any training to use data, so in general we are not really clear – other than what we have taught ourselves – how to use data. That's a useful thing to know because the things I want to talk about are the numbers used in education and what the numbers are used for. Particularly how numbers are used to assess teachers and to assess schools.

So let's have a look at some things which weren't covered in my training as a teacher. One of those things is to look at what numbers actually are and how they can be used. Because much of the use of numbers in education could be termed Cargo Cult Data.

The well-read – or nerdy – amongst you will have spotted a variation on Richard Feynman's term Cargo Cult Science, which refers to activity which has the trappings of real science but which lacks the rigour expected of real science. A huge amount of data in education is Cargo Cult Data.

In the book I say that, "*Just about all of the internal progress tracking data and external test data which is used in English education could actually be more correctly described as cargo cult data – that is to say, it has the appearance of countable, ordinal data which you can torture fairly well - without fulfilling any of the requirements of being statistically valid, approximately measured, error accepting countable data from which one could reasonably draw inferences. The fundamental problem with education data is that it cannot be subjected to close statistical scrutiny.*"

Now, don't get me wrong, numbers are wonderful things. And you can use them to interpret our frequently chaotic world. But using imprecise measure of what children have learned to make judgements about teachers and schools is highly questionable. I approach most use of test data with scepticism bordering on hostility and I suggest that you do the same.

The main trouble with education data is that it simply isn't quantitative data in the sense that most statisticians would recognise and use the term. Most education data is qualitative at best – it's subjective, descriptive and difficult to measure. I don't go too deep into this in the book, but I've been researching a second book I've been asked to write – on Data for Teachers – and it's made me go back to the statistics I studied at university. Yes, I'm an extremely rare primary school teacher in that I've actually done a degree in Economics, Mathematics and Statistics and I've studied in depth how numbers can be generated, manipulated and interpreted. This is, needless to say, highly unusual as far as I can tell.

So here are some things to consider about numbers. Quantative data – i.e. the kind of data I refer to when I say Education Data, that is, data which is based on quantities using a quantifiable measuring process – comes in four main forms:

- Nominal data
- Ordinal data
- Interval data
- Ratio data

Now, these different types of data lend themselves to different levels of manipulation. Nominal Data – Nom meaning name – is simple categorical data. Left/right, county of birth and so on. You can't do much with *Nominal data* as there's no numerical relationship between the categories. You can count things such as languages spoken by a class of children, but you can't really do much with the data. If you can order nominal data, it is then referred to as *Ordinal data* (there's an order). A child who says they *like* maths a lot is clearly different to one who says they like *dislike* maths. A child who gets 19 out of 20 is clearly different to one who gets 1 out of 20 on a test.

But neither of these categories of numerical data lends itself to complex interpretation, for all kinds of reasons which seem self-evident to me but don't seem to be that clear to many of the people I speak to in education.

For example, one of the issues with ordinal test score data is that the distance between the values isn't – for reasons I'll come on to soon – meaningful. All you can say is that one appears to be bigger than another. What you need for higher level analysis – such as using something as simple as a mean – is *interval* data, in which the interval is consistent and meaningful. And for analysis at the highest level – to use complicated statistics – you need ratio data, which has a meaningful zero point.

The easiest way I can think of to explain the difference between interval and ratio is to consider temperature on a Celsius scale. The temperatures of, say, 20 degrees and 10 degrees are clearly 10 degrees apart. But 20 degrees isn't twice as hot as 10 degrees, because zero isn't meaningful – it's just the freezing point of water. In fact, if we use absolute zero - which many of you will know is… -273 degrees C, a change from 10 degrees C and 20 degrees C is actually…

Well, what do you think? Here's a multiple choice question for you. Is the difference between 10 and 20 degrees centigrade an increase of:

a) 300%
b) 100%
c) 30%
d) 3%
e) 0.3%

It's actually a 3.5% increase, not the 100% increase an innumerate person might assume.

So you couldn't really, for example, say that a child who had scored 16 in a test had scored twice someone who had scored 8. You can't use these scores to measure progress. All you can do is rank the children, in rough order of achievement. Unless you have some sense of what progress is *relative to a zero point*, you have no idea how big, or small, the progress made by a child or a class actually is. What's more, since the instruments you use to measure achievement are so variable, you can't even be sure that the data you have is robust enough to be summarised using means, and you certainly can't use the same instrument to test different age groups.

These kinds of problems are what I mean by Cargo Cult Data, and there is an area of research and thinking in which more people working in education should be trained to understand how and why test scores are actually generated: Psychometrics.

What should teachers know about psychometrics?

So, of those who trained as teachers, how many have read or learned much about psychometrics?

I was entirely unaware of even as much as the term until relatively recently. 'Psychometrics' is the name given to the field of study of objective measurement of various aspects of the human mind. A sub-branch of the field is the objective measurement of educational achievement.

Until I finally made time to properly research the best that has been written and said about measuring educational attainment, I suspect that I had a typical teacher's view of the tests which children take. I thought that they were a fairly poor method of assessing what a child had learned, and were clearly biased toward those with, in the jargon, a lot of 'capital', both social and cultural.

As a statistician, I was aware that the numerical result of a test was simply one of a range of possible results which a child might be awarded, and that there are a number of factors which affect the result of any test. But I had not really considered what a test was measuring and why.

Having read a lot of educational research written by those who design tests, I have found out a lot which simply had not occurred to me. I suspect quite a bit hasn't occurred to many teachers and policy makers, either. For example, it hadn't really occurred to me that tests are designed to sample knowledge, skills and understanding to provide an estimate of the full range of knowledge, skills and understanding (the 'domain') a child might possibly have.

Given what happened in the recent General Election, a good analogy would with political polls and the final election result. The polls can only hope to use the responses of a sample of the population to anticipate the response of an entire population. And sometimes they can be spectacularly wrong.

The following, from Daniel Koretz's Measuring Up, published in 2008, summarises what tests are designed to do:

'*The results of an achievement test – the behaviour of students in answering a small sample of questions – is used to estimate how students would perform across the entire domain if we were able to measure it directly.*'

But of course, creating these samples is incredibly difficult. It simply isn't possible to test everything which you might want to test. Even if the test is very good at assessing achievement, there may be other aspects of school quality which a test simply can't tell you, such as whether the children have enjoyed the experience of learning the subject, or whether they have come to hate it with a passion.

A further problem with using tests to judge schools is that this in itself can cause the sampling to be skewed. And if the samples are skewed – as per the general election polling - the estimates are meaningless.

As soon as you understand that tests are simply samples which provide estimates, one of the problems inherent in using test results to judge schools and teaching becomes more clear. As Daniel Koretz notes, "*A failure to grasp this principle is at the root of widespread misunderstandings of test scores.*"

There are many things which flow from this principle. One is that there may be things which all children learn, but which never appear on the test because such questions aren't actually useful in a test designed to rank children. Secondly, teaching to a test distorts what the test is designed to do. If a child has learnt a method to gain marks on the test rather than further their understanding of the subject, the test loses a great deal of its validity. This is also true if an area which never appears on the test is not taught. It can be argued that any teaching which focuses on the test rather than the subject distorts the result. And whether we like it or not, there is a huge amount of teaching to the test in our schools.

In David's book, I used ideas put forward by teacher, physicist and psychometrician Noel Wilson who has set out the many issues which have been demonstrated by Psychometricians to actively prevent anyone from objectively measuring educational knowledge with any degree of accuracy.

In brief, any measurement of anything which is continuous and equally spaced (such as time or length) has to be measured using a specified standard unit, and this will necessarily involve a degree of error. The units have to be defined as a standard, and the standard has to be completely accurate by definition. So a degree centigrade is a degree centigrade because we say so; it is a definition and not a measurement.

Additionally, when measuring a temperature, for example, a measuring tool such as a mercury thermometer, a gas thermometer, a pyrometer or a Langmuir probe – some kind of instrument – has to be used. Whatever the instrument measures will introduce an element of error.

In attempting to measure educational knowledge, the standards are not standard and the measurements are not accurate. What is the 'unit of education'? There are other errors which enter the fray. The interference effect says that 'any measuring instrument distorts the field it is intended to measure', and it makes measuring something as well understood as temperature quite difficult. It also means that the specific content of any test designed to assess knowledge affects the mark anyone taking the test is awarded. Uncontrollable boundary conditions mean that humans respond to tests in unimagined ways.

Giving the test is an artificial situation and is not a true indicator of knowledge, just as a wind tunnel test of a car cannot accurately test the actual performance on any given road on any given day or in any set of weather conditions: "Perception and conception, and hence response, to 'identical' situations invariably differ, as the variables that affect such reactions – attention, mood, focus, metabolic rate, tiredness, visualisations, imagination, memory, habit, divergence, growth etc. – come into play.

Now, as I say, this is a well-established field of research and yet those in school seem to be almost entirely unaware of the problems which have been identified with trying to use test scores to assess what a child knows and what they have learned. For example, Lord Bew noted in his 2011 report on Key Stage 2 testing:

"*It is generally accepted that any test or examination, however well-constructed, will always include a degree of measurement error. We understand that, as with all tests where pupils are categorised, the level thresholds in Key Stage 2 tests mean that one mark can make the difference between one level and the next. That mark could be lost or gained through a pupil mis-reading an instruction in*

the test or making a fortunate choice in a multiple-choice question, or through slight variations in marking practice. These differences will be highly significant for the individual pupil."

Lord Bew also noted that Dylan Wiliam has suggested that 32 per cent of pupils could be given the wrong national curriculum level. Dylan observed, "*we must be aware that the results of even the best tests can be wildly inaccurate for individual pupils, and that high-stakes decisions should never be based on the results of individual tests*".

What should teachers know about education data?

Well, firstly, you need to know that any data based on test scores is very fuzzy indeed. Tests are designed to rank children, and it's extremely hard if not impossible to summarise a child's learning in a number, especially in a point estimate. Assessments are clearly fuzzy, and the child previously referred to as 'working at level 3a' is probably not 'working at level 3a'. It's worth remembering that levels were designed to help teachers teach in a rough sequence, not to attach numbers to children. Secondly, any system which attempts to reduce learning to a number is fundamentally flawed. That's why we use grades in England, and not numbered scores; the grades are inherently fuzzy, as they should be. All you really know is that someone awarded an A is probably different from someone awarded a D.

Teachers should also know that Ofsted grades, which are based on extremely fuzzy data, are Not Even Wrong. I like this phrase a lot. Not Even Wrong is a phrase attributed to theoretical physicist Wolfgang Pauli (1900-1958). "This isn't right," he is supposed to have said of a student's physics paper. "It's not even wrong."

Ofsted Inspectors use data presented in RAISEonline prior to entering a school. Despite what the inspectorate claims, the vast majority of judgements are solely based on data, as is obvious when the correlation between Overall Grade, Quality of Teaching and Achievement of Pupils is. Firstly, most inspectors clearly do not understand the data with which they are presented, and even if they did, the central Ofsted machine dictates that 'good' data indicates a 'good' school. Secondly, school data is analysed centrally using statistical procedures designed to extrapolate from a sample to a population. But because schools are not randomly selected samples of the population, this kind of analysis is invalid, and thus any conclusion will give a false impression of some schools being 'good' and others 'bad'. Finally, because school inspectors have access to this dubious data prior to entering schools to inspect them, they cannot help but be biased by their preconceived notions of the school's effectiveness; their judgements thus suffer from the halo effect (which David explains in his book) and this influences many of the assessments inspectors make. In England and Wales, all school data is presented centrally using RAISEonline.

Because this uses standard statistical techniques, which assume that a given sample is truly representative of a wider population, this gives the RAISEonline data analysis a spurious sense of validity. The use of standard statistical techniques – calculating statistical significance and confidence intervals to 'test' how a school compares to the national population – is not valid for a number of reasons. Firstly, no given school can be said to be drawn from the wider population because families and their children are clustered geographically and demographically. Secondly, even if the data could be regarded as a sample of the population, it isn't valid to treat it as such. It would have to be regarded as a clustered subsection of the complete population data which has been collected, and if that were the case, statistical techniques for sample data simply aren't valid.

Despite these obvious problems, RAISEonline sails merrily on, colouring certain items of data green and blue, comparing school data to national means as if it were a random sample of an unknown population and generally creating a completely misleading impression of any given school.

When we make assumptions based on such data, we delude ourselves. So, given the inaccuracy of assessment 'data', and the invalid analysis it presents, there is simply no point in drawing any worthwhile conclusions from RAISEonline.

To conclude on *'What should you know about education data?'*, I'd like you consider the average monthly temperature here in Crowthorne. Specifically, I'd like you to consider the average temperature in June. It's a nice day today. Have a guess what the maximum temperature was in June 2013. It was measured by the Met Office as 20.3^{o} C 26km away at Heathrow (which is the closest Met Office weather station to Crowthorne).

How about in the years before 2013? Well, here is the maximum and minimum temperature recorded by the Met Office, along with rain in mm and hours of sun.

So, here they are:

Heathrow

Year	Max	Min	Rainfall	Sunshine Hours
2013	20.3	11.2	11.6	157.5
2012	19.4	11.6	110.8	118.5
2011	20.7	11.0	84.0	173.5
2010	23.5	12.1	12.4	220.1
2009	22.4	12.2	34.0	192.8
2008	20.8	11.9	45.6	201.9

Of course, all of these are ratio data, even if sunshine hours is a fairly fuzzy measure. And you can see the variation. If you had no understanding of temperature, rainfall and sunshine you could see that those in charge of the rain dance in 2013 were in trouble, and those in charge of sunshine in the same were clearly brilliant at their job.

Compare that to the results of a typical primary school. I've called this on Seaside Primary, but it's a real school in North Yorkshire.

This is a fairly typical two form entry school. These are the percentages of children achieving level 4 or above in reading, writing and maths:

2013: 77%

2012: 70%

2011: 58%

2010: 69%

2009: 77%

2008: 76%

As you can see, there is no pattern. Unless a school consistently records 100 per cent, there never is a pattern for any school, in any historical data. And neither should there be. This is because the data is based on children's results, and children are complicated and individual, and the school population in any given school is statistically too small to make meaningful generalisations.

The results vary just like the weather, which is also notoriously hard to predict and prone to superstition and cognitive biases. Of course, you can find a mean, and a standard deviation, and you could expect the weather in any given year to be a point estimate within the distribution implied by the mean and standard deviation – but any given year will be random within that distribution.

As any financial advertisement will tell you in the small print, past performance is no guarantee of future success. Long-term trends in something as complex as educational outcomes are – unless you mess with the data by, you know, making the tests easier, selecting by ability or dis-applying certain children from assessment, or simply not reporting stuff – always random.

And that's what you should know about education data. The noise pretty much overwhelms any signal you might try to identify if you are using when you analyse test score data and treat it as ratio data. Test scores are imprecise, badly measured and random by nature.

What should we be doing?

The truth is, not very much. Children's learning and progress is too complicated to be reduced to a few numbers. While tests only assess test taking, in an extremely fuzzy way, it is possible to track the acquisition of certain kinds of fundamental knowledge which is vital for pupils to make progress in education. For example, their knowledge of number bonds and times tables, since this is simply testing recall of factual information which is clear and unambiguous.

It might also be worth keeping an eye on pupils' knowledge of the alphabet, phonemes and graphemes, and punctuation marks for those who have yet to master writing. It's hard to track such progress in a meaningful numerical way, let alone trying to measure those things which are much more complicated core concepts. However, monitoring progress is possible, and teachers can track whether pupils are using certain key concepts correctly. But this does not mean it is possible to assign numbers to this in the way that most schools currently employ.

Beyond this, it's clearly possible to record indicators of children's progress against what you think they should have learned based on your teaching. My guess is that, currently, this is what most teachers actually do, even if the data then ends up being used as cargo cult data. For example England's national curriculum levels can be quite useful as indicators. Yes, using levels is hugely subjective and susceptible to all kinds of biases, and we almost certainly under-record some children and over-record others.

As long as no one tries to abuse this information with the widely misunderstood techniques often used to summarise actual data (a forlorn hope in most cases), I have no problem with it, and find it useful.

Age within a cohort is highly important but is completely ignored by most schools. This should be tracked, as it is actual data – a child's age is a very good, robust measurement – and it does affect what a child can do. Too often, high ability groups are simply the older children in a class, with lower ability groups made up of the younger children. The middle of the school year – 2 March – should be central to every teacher's understanding of the relative abilities of the children in a class. This is data worth using.

Maybe I am wrong to feel that schools are judged unfairly by the use of dubious made up data. It could be that questioning the other pressures on English education are more important, and the issues being raised by those in and around education in our more connected world need to be dealt with first. Misuse of Cargo Cult Data could be halfway up the flawed foundations of the way we look at education; there could be more fundamental issue to address.

But for now, I hope that more teachers look behind the numbers and consider the question: *What If Everything You Know About Education Data is Wrong?*