Understanding the quality of data: a concept map for 'the thinking behind the doing' in scientific practice

Abstract

Recent school science curriculum developments in many countries emphasise that scientists derive evidence for their claims through different approaches; that such practices are bound up with disciplinary knowledge; and that the quality of data should be appreciated. This position paper presents an understanding of the validity of data as a set of conceptual relationships, illustrating the application of the network of ideas and their inter-relationships necessary for the 'thinking behind the doing' with examples from practice. We explore ways in which this understanding of data is inherently related to underpinning disciplinary ideas. We suggest how the recognition of a conceptual basis for understanding the quality of data represents an ontological shift with respect to widespread characterisations of scientific practices which addresses some long-standing issues in science education research, policy, curricula and practice.

Introduction

In many countries science curricula now represent science as including not only 'the products' of science (the substantive facts, theories and laws, sometimes referred to as the content knowledge (e.g., Organisation for Economic Co-operation and Development (OECD), 2013)) but also 'the processes and characteristics of the scientific enterprise' (Roberts, 2011, p. 12). These developments all share the broad aim of Scientific Literacy (Bybee, 1997; DeBoer, 2000; Laugksch, 2000; Roberts, 1982, 2007). Despite being 'a rather polysemic expression' (Martin, 2011, p. 90) a common feature is to 'understand the methods by which science derives the evidence for the claims made by scientists, [and] to appreciate the strengths and limits of scientific evidence' (Millar & Osborne, 1998, p. 2004). Large-scale international science assessments such as PISA (OECD, 2013) and TIMSS (Jones, Wheeler & Centurino, 2013) also reflect this curriculum emphasis (Kind, 2013a). Since DeBoer (2011, p. 569) suggests that 'in some countries … the approach is to match educational programs to the framework that is guiding the development of the

international assessments' it is reasonable to assume that attention to 'the doing of science' as well as the traditional substantive content substantive will become even more widespread.

Over the years, how this 'doing of science' has been conceived and expressed in the research, policy and assessment literature has differed (Hofstein & Lunetta, 2004; Jenkins, 2009; Kind, 2013a; OECD, 2013). For many decades, it was conceived in terms of various 'processes' to be acquired through practice (Millar & Driver, 1987). However, recognition that the 'doing' is 'supported by the *integration* of science concepts and processes, metacognitive processes, critical reasoning skills, and cultural aspects of science' (Cavagnetto, 2010, p. 337; emphasis added) has come to recent prominence. Many researchers (e.g., Lederman, et al., 2014; Lubben, Sadeck, Scholtz, & Braund, 2010; Roberts & Gott, 2010; Schalk, van der Schee, & Boersma, 2013; Tytler, 2007) have moved beyond describing what scientists do (wherein any understanding may be implicit) and explicitly articulate some of the ideas required to *understand* evidence since:

> At the core, science is fundamentally about establishing lines of evidence and using the evidence to develop and refine explanations using theories, models, hypotheses, measurements, and observations. (National Research Council, [NRC], 2007, p. 18)

The *PISA 2015 Draft Science Framework* (OECD, 2013) addresses the importance of evidence in both its 'procedural knowledge' and 'epistemic knowledge' elements and argues that PISA's 2015 definition [of scientific literacy] represents 'a more detailed specification of particular aspects that were embedded or assumed in earlier definitions' (OECD, 2013, p. 10), thus recognising the importance of making the implicit explicit and clarifying the construct for assessment (Wiliam, 2010).

In the US, the new *Framework for K-12 Science Education* (NRC, 2012) the dimension of 'Scientific and Engineering practices' corresponds to understanding evidence. The document states that 'engaging in scientific investigation requires not only skill but also *knowledge* that is specific to each practice' (p. 31; emphasis added). Other recent developments (including Achieve, Inc., 2013; Australian Curriculum, Assessment and Reporting Authority, 2012; Department for Education, [DfE], 2014) and science education research (e.g., Lederman et al., 2014) give greater emphasis to

students being able to understand the diversity of scientific empirical practice; the important relationship between substantive knowledge and this 'doing' aspect; and the importance of students being able to use their understanding to evaluate empirical work and reason with evidence as well as being able to carry out practical work.

Evaluating empirical evidence requires understanding about the validity of data. This understanding represents the 'thinking behind the doing' of science. In 1994, Millar, Lubben, Gott and Duggan called these 'concepts of evidence' and subsequently a knowledge base to develop an understanding of empirical evidence has been specified (Gott, Duggan, Roberts and Hussain, n.d.).

Viewing scientific practice as a conceptual knowledge base to be understood rather than skills or processes to be acquired represents an ontological shift in its characterisation. While recent curriculum documents reflect this shift, since scientific practice is concerned with 'doing', curriculum specifications tend to describe the processes involved and the conceptual basis - important for teaching and assessment - is not always so explicit. Furthermore, since the process view and its association with a single scientific method are so deep rooted in customary school science, as shown by Abrahams & Millar (2008), there is the danger that the significance of the conceptual shift inherent in the new curriculum documents could be lost if curriculum developers, awarding bodies and teachers interpret them from a process perspective.

Concept maps (Novak & Gowin, 1984) are widely used to represent substantive understanding in science (e.g., Johnson & Papageorgiou, 2010; Kinchin, 2010; Shymansky et al., 1997). Equally, if 'the thinking behind the doing' is also a knowledge base of concepts to be understood (rather than 'processes' to be routinely mastered) it ought to be possible to represent that understanding with a concept map (Novak, 2010). A concept map should help to emphasise the difference between an underpinning 'conceptual' characterisation of scientific practice necessary for meaningful learning (Novak & Cañas, 2007) and a surface description of processes. This article presents an attempt to produce a concept map for understanding the validity of data.

The map is informed by the concepts of evidence and their inter-relationships (Gott, Duggan, Roberts and Hussain, n.d.). The current version has been derived from more than a decade of our, and colleagues', research, teaching and assessment. It has been

through many iterations following discussions at both professional and academic conferences and has recently been informed by ideas from Cañas, Novak and Reiska (2015) with the aim of it having explanatory power, being clear and concise and having balance (Kinchin, 2015).

Our map (Figure 1) centralises the question of the concept of **validity of data** since the degree of confidence in the validity gives it weight as empirical evidence for a claim. The network of ideas drawn on in judging the validity of data represents an understanding about the quality of data. This understanding is inextricably linked with substantive understanding and those concepts directly informed by substantive knowledge are highlighted with a shadow on the box. To reiterate, by substantive knowledge we include 'facts', laws, models and theories of the disciplinary sciences. We would suggest that the map as a whole encompasses most if not all of 'scientific practice' in schools.

<<INSERT FIGURE 1 ABOUT HERE>>

We have not seen the understanding of evidence being addressed through a concept map in the literature and believe this offers a valuable perspective for science education. The map of the conceptual relationships helps to highlight important points that are not always so easy to communicate to a wider audience more attuned to a process view. As a concept map, it is important to distinguish it from a flow diagram; it does not represent any particular procedure or approach. The map is structured from the perspective of carrying out an investigation i.e. data is being collected to answer a question.

In this article, firstly, we will explain the meaning of the terms and propositions in our map, using examples to illustrate how the ideas and their relationships are applied in the 'thinking behind the doing'. The map does not represent the processes but shows the thinking – the ideas and their inter-relationships - behind decisions; so we give examples from different areas of science to exemplify the application of the ideas as they inform practice. Since the arrows are not processes, but are conceptual links, there is not necessarily any one starting point or direction of travel in practice. However, for convenience in our explanation we will work systematically through the map. After examining its detail we go on to consider some important general points which emerge from the map as a whole with respect to the nature of science, the

notion of ‘scientific method’ and the relationship between substantive theory and scientific practice. Finally we address how the ontological shift in the characterisation of scientific practice from routine process to conceptual understanding has implications for practical work in schools.

A concept map for understanding the validity of data

We will explain our map with particular reference to two contrasting investigations; called Springboard (Figure 2) and Shrimp (Figure 3). These investigations allow us to discuss the understanding represented by the whole map. We recognise that some scientific practices may only focus on specific regions.

<<INSERT FIGs 2 + 3 ABOUT HERE>>

Broadly speaking, Figure 1 has two inter-related sides. On the left is thinking about variables and on the right thinking about measurement. We will start with variables.

Variables

All investigations involve **defined variables**. These variables are the creation of existing **substantive knowledge** - they constitute the disciplines biology, chemistry, physics, earth science etc. Scientific observation is dependent on ‘seeing the world’ through science’s ‘conceptual spectacles’. As Lederman et al. (2014, p. 68) state, investigators ‘need to have specific knowledge that has been melded into some curious pattern or question’. Any limitations in understanding of pertinent substantive ideas affects what can be observed (Haigh, France, & Gounder, 2012).

Variables can be categoric, where the value is descriptive (e.g. a material or species) or continuous, where the **value** lies on a numeric scale. (e.g. length, temperature).

Although rooted in the scientific conceptualisation of our world, the variables of Springboard do not draw on specialised knowledge. With regard to load, there is no need to distinguish between mass and weight, and distance is part of everyday thinking. Explanations may use ideas relating to force and motion, but that is another matter (see later). Shrimp draws more directly on scientific knowledge. It involves the identification of freshwater shrimp *Gammarus pulex* (a genus of freshwater amphipod crustacean) and other variables such as water speed, pH and oxygen concentration.

Variables may have a **relationship** with each other. For Springboard we could ask how the height reached by the toy figure is related to the load. For Shrimp we could wonder where shrimp like to live. The decision to look for a relationship derives from existing knowledge. With Springboard, a hunch based on everyday play experiences may be the basis for the question, though the expectation of a specific relationship is of itself scientific. Shrimp stems from the substantive idea of different species requiring different ecological conditions.

The supposition of a relationship implies that a change in one variable corresponds to change in the other. For convenience the two variables in focus can be called the **independent** (IV) and **dependent variables** (DV) respectively[i]. As will be discussed later, the assignment of the DV and IV labels does not necessarily presume a causal link. However, **confounding variables** are other variables which are thought to affect the DV. So, the identification of confounding variables absolutely draws directly on substantive knowledge and is limited by that knowledge. There has to be a reason for deciding upon a particular variable, if only because it might be relevant. Confounding variables must be **controlled** in some way to isolate any relationship between the chosen IV and DV.

For a **continuous** IV decisions about the **range** and **interval** of their **values** have to be made. The **range** is crucial in capturing the full picture of any relationship. Substantive theory can inform the thinking, but trial runs will be needed to explore a suitable range in the circumstances. **Interval** can also be important, especially where closer readings are helpful in picking up any maximum, minimum or inflection points. Decisions about **intervals** of the IV can also be made iteratively, in response to the data gathered, with an eye on any emergent pattern. Again, substantive thinking can lead to the anticipation of such eventualities. The idea of range can also apply to a categoric independent variable; for example, in a question exploring a property of materials one would need to decide how many different materials to test.

For Springboard, height reached is the DV and load is the IV. **Confounding variables** include: material of ruler, dimensions of ruler, overhang, clamping position, position of figure on ruler, position of mass, mass of figure, dimensions and shape of figure, mass of string, temperature, air movements. All of these confounding variables can be directly **manipulated** and each fixed at a certain value. Some are more important than others. The length of overhang is a critical variable while it might be

assumed that any small fluctuations in temperature will have negligible effect and might not be given special attention. For those variables to be fixed, what should their values be? Decisions about one cannot be made without reference to others and determining their impact on the DV. The values need to be chosen appropriately. Too small an overhang and/or too heavy a toy figure may not produce a measurable height reached for the range of available loads. Trials to establish a range for the IV must be in conjunction with exploring fixed values for the CVs in relation to the impact on the DV. In some contexts, a confounding variable is manipulated not by fixing to a certain value, but by ensuring any fluctuations are the same across different values of the IV. This is often seen in the setting up of a 'control' in biological contexts.

For Shrimp, the number of shrimps in an area is the DV. Variables which might affect the DV would include: pollution, aquatic vegetation, pH, velocity, oxygen in water, substrate type, substrate size, depth, surrounding and upstream vegetation, other animals, time of day and time of year. While selection of time of day and time of year are under the control of the investigator, for the others in the natural context it is not possible to isolate one variable from another and the values cannot be directly manipulated. (This could only be done by lab-based modelling of the stream; as illustrated for ponds by Lehrer, Schauble, & Lucas, 2008). Data from variables that cannot be manipulated are sometimes referred to as observational data (e.g., Gray, 2014). Identifying all relevant variables depends on substantive knowledge as does the recognition that some of these variables may be **co-variables** (i.e. oxygen concentration and substrate size are both related to velocity).

In a survey like Shrimp the approach is to collect data on all of the variables at as many different sites as seems reasonable (based on substantive understanding) or on some systematic basis or as is possible in practice. P*ost-hoc* comparisons can then be made. Hodson (2014) points out, that such 'data mining' is now more common across experimental science. Firstly, co-variation can be examined and if confirmed these can be treated 'as one'. Different variables can then be considered in turn as the IV with the others as confounding variables being **matched** by **selection** of sites where values are similar. For instance, the relationship between velocity and shrimp numbers could be determined at sites matched by selection of, for example, the 'absent' values for the variables pollution and aquatic vegetation i.e. at sites where there was no pollution and no macroscopic plants growing.

The naturally occurring values of any **independent variable** have to be chosen (rather than actively changed) to capture the natural **range** of any IV considered in the stream and the range would need to be considered too in relation to the **magnitude of any effect** which will become clearer during data collection. Site selections for the values of the IV could be random, or they could be selected and consider as stratified (encompassing the range of velocities, say). As measurements are taken throughout the stream, the investigator would have to consider whether enough intervals for any variable considered as an IV had been measured to identify any potential pattern.

Measurement

Variables have **values,** categoric or continuous, all of which require **measurement** by an **instrument** of some kind. For continuous variables, apart from a ruler to measure distance, most if not all instruments rely on previously established relationships between variables and are therefore products of existing substantive knowledge. For example, common thermometers presume a linear thermal expansion of mercury or alcohol.

All measuring instruments have a built-in **degree of uncertainty –** the recorded values of all variables are approximations. Foremost, in a school context, when measuring continuous values is the resolution of the scale. In the case of a ruler marked out in one mm intervals, it is possible to read to the nearest 0.5 mm (or even smaller) with high confidence. For a digital instrument such as a top-pan balance giving a readout to the nearest 0.1 g, for a value recorded as 5.0 g we can only be sure we have a mass somewhere between 4.95 g and 5.05 g. High level work would need to take the detailed specification of an instrument into account. For example, a thermometer has an error associated with the consistency of the bore's cross-sectional area and a certain depth of immersion is stipulated. Ideally, one chooses an instrument where the magnitude of the uncertainty is small in proportion to the value being measured.

Our definition of an **instrument** also includes the totality of how it is used to measure a particular variable. For Springboard one instrument for measuring height reached could be 'using a metre rule and sighting by eye' and another could be 'a metre rule and recording by video'. The former would have a greater degree of uncertainty associated with judging the highest point of the trajectory. In Shrimp, the instrument

for measuring the number of shrimp is 'kick sampling'[ii]. The number of netted shrimp can be carefully counted, but there will be uncertainty about the precise area disturbed, the proportion of shrimp dislodged and the proportion of those actually caught in the net.

An instrument, then, can only measure a continuous variable to a region (not a point) and its **accuracy** relates to how close that region is to the true value. For an accurate measurement, we would expect the true value to lie within the region. A systematic error can arise if the scale is not calibrated correctly.

Sometimes it is not possible to measure a variable directly and a proxy is used instead[iii]. For Shrimp, the most relevant water speed will be that at the bottom. However, without very sophisticated instruments this will be difficult to measure. Instead, a pragmatic decision may be made to measure 'water speed at the surface'. How well this substitute stands in will be open to debate.

The concept of **measurement** (although often referred to as 'observation'; see, for instance Gray, 2014) can also be applied to a categoric variable, where qualitative descriptions are the values. Here, the measurement entails the recognition of the defining features of the variable, with the substantively-informed discernment of the observer acting like an instrument. Our Shrimp investigation rests on the capacity to distinguish *G. pulex* from other creatures. There may be uncertainty with some specimens. Placing vegetation into different categories will also require judgment[iv].

Reliability of the DV

With a single measurement of the DV, we only know an approximate value of the DV for what will also be approximate values of the IV and CVs. In many cases, this is all that is possible and the quality of the data and patterns will be judged with recourse to the ideas below but without being able to apply them empirically. Where the other variables are readily manipulated, the **reliability** of the DV can be assessed by repeated 'takes'. Here, the degree of uncertainty associated with each instrument of the investigation makes a combined impact on each value of the DV measured. In setting up for each take, the starting conditions will only be the same within the limits of the instruments and differences may result in a different value of the DV being recorded (subject to the uncertainty of the DV instrument).

For Springboard, the important variables such as overhang, load and toy figure can be left unchanged between 'takes' and perhaps only the position of the toy figure might vary slightly giving differences in trajectory. Other experiments may require more re-measuring out of quantities in setting up each time.

However, the reliability of the DV will also depend on what we call the **repeatability of the event**. By this we mean the extent to which all other confounding variables have been identified and controlled. For Springboard, we may not have thought about the cutting of the string. It will be important to have a sharp pair of scissors since any snagging or pulling on the string will affect how bent the ruler is on point of release. Although identified, we may not have been able to control for air currents. As noted earlier, the identification of confounding variables is as good as our substantive understanding of the situation.

Increasing the **number of repeats**, if possible, narrows down the region where we can say the true value of the DV lies (for each value of the IV in a region according to its measurement). Statistically speaking, increasing the number of repeats reduces the standard deviation of the mean, otherwise known as standard error ($SE = SD/\sqrt{n}$). We can be 68% confident that the true value lies in the region between one SE either side of the experimental mean value. Extending the region to two SE either side of the experimental mean gives a 95% probability of covering the location of the true value.

What constitutes a good enough reliability is a matter of judgment in the circumstances. The key issue here is how the SE of the DV compares to the magnitude of any change over the range of the IV, since this will determine the extent to which any intrinsic relationship can emerge[v]. Trial runs to get a sense of the **variation in the data** and the **magnitude of the effect of changing the IV** are essential to the planning stage of an investigation. If the variation (SD) is very small in comparison to the change with IV then relatively few repeats are necessary to give a good enough SE. (With a few repeats its computation would not be appropriate, in practice.) If the variation (SD) is large in comparison to the effect of changing the IV one should first think about ways of reducing the variation - better instruments and/or thinking again about possible confounding variables and their control. Failing that, if feasible, the number of repeats will need to be increased until the SE is small enough to reach a conclusion. Substantive understanding of the various factors will inform the best course of action. For Springboard, the uncertainty in judging the highest point by

eye will most likely be the biggest contributor to the variation and the use of video technology would be worthwhile. If not available, since the time for a take is relatively short, a large number of repeats is feasible.

We have focussed on the reliability of the DV, but the need for **repeated measurements** can also apply to other variables. For example, in Shrimp, the velocity at the surface can be measured by timing a floating object (such as a ball) travelling a fixed distance downstream, such as one metre. However, since the movement of the object will be affected by inconsistencies of wind and eddies a number of repeats will be necessary to establish the variation and a good enough mean for comparison with other sites.

Repeated measurements of a variable are not always practicable or even possible and any assumptions must always be acknowledged in the interpretation of the data. The disruptive kick sampling in Shrimp prohibits repeated readings of the DV in the same spot on the same occasion. One way round this would be to sample in adjacent sites where conditions are similar. Another would be to return at a later date after the stream has had the chance to settle down (if the different time of year was not an issue). Collecting data from other streams (with similar key characteristics) would also be useful. The aim is to collect as much relevant data as possible (subject to ethical constraints). By having many sites the intervals of all IVs would most likely be smaller, which would also help to better establish any pattern in the data. As with shrimp numbers and water velocity, if the effect of changing the IV is greater than the variation, a scatter plot will indicate a trend and the strength in the relationship can be quantified through the use of statistics such as the correlation coefficient and its probability.

Before completing our consideration of reliability, we must mention ‘human error’ which for reasons given below, is excluded from the map.

Human Error

We have noted the constraints of our sense organs when incorporated into an instrument (e.g. judging highest point of flight or accurate identification of specimens). However, that people can be sloppy and make mistakes is self-evident. That there is an inherent variation in the measurement of a DV, no matter how carefully done, is the crucial point to understand. It can then be appreciated how very

tightly manipulated contexts can lead to little variation (or no variation that can be picked up by our instruments). Too much emphasis on 'anomalous' readings arising from 'human error' carries the danger of offering an easy distraction – sliding to all variation being put down to human error (Fairbrother & Hackling, 1997).

Variables that are objects

Returning to the left side of the map, where an object is involved in an investigation, we must consider the variation amongst objects of that type. For living things, sizeable variation between individuals in their behaviour is to be expected and needs to be considered. In our Shrimp survey, we assume that we find enough shrimp to have a **representative sample** of the species. A reason to count the shrimp at a large number of sites is to increase the number sampled. In a lab-based modelling of the stream, we would need to decide on a number of shrimp to use – we could not go on the behaviour of one shrimp. Similarly, an investigation to establish *the effect of different fertilizers on the yield of tomatoes* will need more than a few tomato plants. What constitutes a good enough **number of specimens (sample size)** will depend on the **variation in the data** in comparison to the **magnitude of the effect on changing the IV** (as with deciding on the number of DV repeats). For the lab-based stream model, the variation relates to how the numbers at each site change as shrimp move about. For the tomato investigation, the variation will be in the yield per plant.

Social science and medical studies will select groups through a screening process on confounding variables. If a 'male middle class smoker and drinker' is put in the treatment group, then a **matched** 'male middle class smoker and drinker' is also put in the comparison (control) group. Of course, what counts as 'male', 'middle class', 'smoker' and 'drinker' will need to be defined. If sample groups are large enough one can move into the territory of a randomised controlled trial (RCT). Subjects are assigned to treatment groups by a **random** process. With a large enough sample size it can be assumed the multifarious confounding variables will even out so the only difference overall is the treatment applied to one group and not the other.

For objects manufactured under tight quality control we expect little variation in their characteristics and a small sample, if not just one specimen, is often assumed to be representative. In the case of Springboard, presuming no significant wear and tear, we would expect any relationship found for one ruler to apply to all rulers of that type.

Our investigation would not involve testing lots of the same type of ruler. In chemistry, the objects are 'substances' where we expect no variation within a type but where the purity of a sample is an important issue.

Peer review

We have introduced our map from the perspective of the investigator. The ideas relating to variables and measurement are equally important from the perspective of others. The first consideration in the evaluation of any claim, be that by peer review (both before and after publication) or, less formally, by any sceptical person, is the validity of any presented data. Sufficient information must be known about the circumstances with respect to variables under which the data were generated and the reliability of measurements in order to judge the quality of the data and the limits of a claim. In the UK, the Royal Society's motto '*"Nullius in verba"* which roughly translates as "take nobody's word for it" … is an expression of the determination … to verify all statements by an appeal to facts determined by experiment' [vi](n.d.). The specific claim from any empirical work should reflect the decisions made in the application of the understanding shown in the map and the confidence in the pattern.

Once data are judged good enough to reveal a distinct pattern, there is then the issue of its interpretation. Explanations and the appropriateness of any generalisations beyond the data will draw on existing knowledge and theory. The relationship between evidence and explanation takes us into the complex area of the philosophy of science which our map does not expand upon. In the next section we confine ourselves to a few key points relevant to school science that ensue from linking evidence back to existing substantive theory

Discussion

As noted earlier, we suggest that the recognition of a conceptual basis for understanding the quality of data represents an ontological shift with respect to widespread characterisations of scientific practices. We would argue that this has implications for addressing some long-standing issues in science education research, policy, curricula and practice.

Our map focuses on the *ideas* involved in carrying out a scientific investigation or inquiry from initial observation to judging the quality of the data, and firstly we turn

to how this understanding underpins the resultant claim and its position in the broader substantive theory. This is important for an understanding about the nature of science (NOS).

Evidence and explanation

Causality

Where a relationship is found, it could be causal, an association or pure chance. Acceptance of causality depends on there being a mechanism from established substantive theory that explains how one isolated variable could cause the change in another. In the case of the Springboard, there is little doubt that the ruler is directly causing the figure to fly upwards and that a greater distortion (load) gives a greater height reached. In Shrimp, the distribution of *Gammarus* does change with the water velocity but the latter is not thought to be directly caused by the former. Here, the explanation relates to the co-variation of velocity with dissolved oxygen concentrations, substrate size and available food. The relationship between water velocity and shrimp numbers is an association. The almost perfect correlation between margarine consumption and divorce rate in the US state of Maine (Fletcher, 2014) would appear to arise by pure chance.

Prediction and retrodiction

Perhaps the defining characteristic of scientific theory is the way it responds to new evidence and moreover how it enables predictions (hypotheses) to be made, which in turn may be tested by experimentation. If valid (as far as can be judged) data do not agree with the prediction, some accommodation in the theory is needed, which could range from amendment to abandonment. In the case with some examples from the 'historical sciences' (Gray, 2014) where testing is not possible (other than by modelling), the 'prediction' or 'retrodiction' can be about other instances of the phenomenon that has already occurred or other lines of evidence:

> *Retrodiction* … is the process of inferring the past from the present... Darwin, for example, retrodicted that many intermediate forms of life would be found in the fossil record … Similarly, cosmologists were able to retrodict from the big bang theory the existence of cosmic microwave background radiation (Gray, 2014, p. 333).

Competing theories

Sometimes the connection between variables can be determined empirically but the relationship could be explained by very different theories. For instance, that the time for a candle to extinguish within an upturned jar depends on the volume of contained air is readily demonstrated. However, in the late 18th Century, proponents of the established phlogiston theory and the newly proposed oxygen theory disputed the explanation. The eventual acceptance of the oxygen theory occurred after a body of evidence, from different empirical works, accumulated and, following judgement by peers, was considered to be the most parsimonious explanation consistent with the evidence. Here the combustion of hydrogen to give water as the only product was decisive (Conant, 1957). Scientific theories are tentative and underdetermined by the evidence (Jimenez-Aleixandre, Rodriguez & Duschl, 2000; Lederman et al., 2014).

Overall, it is important to appreciate that theory is not evidence. However, a person's theoretical persuasion has a strong influence on what is sought and accepted as evidence. There is the danger that poor data are accepted uncritically because they fit with expectations and sound data are dismissed on superficial grounds because they don't. Such influences are at play in peer review and political circles, as well as in science classrooms (which we discuss further below).

There is no single scientific method

The concept map represents an understanding of the validity of data and contains no ordered series of procedures or processes. It is not a flow diagram of the sort that is often associated with descriptions of scientific 'chains of practice' (Kinchin & Hay, 2007). The arrow directions in conjunction with the linking terms are there to represent the propositional relationships which give meaning to the concepts and do not imply a procedural sequence. To reiterate, this is a map of the 'thinking behind the doing', whatever form the 'doing' might take (Roberts & Gott, 2003). The conceptual overview represents a network of intricately linked ideas, and decisions when investigating are based on nuanced application of these ideas, involving mental juggling as juxtapositions and contingencies are considered according to context. In terms of validity, there is no distinction between approaches (such as an 'experimental approach' or an 'observational approach') to finding patterns in data (Cleland, 2002). No one approach is privileged over another; the key issue is what is appropriate

depending on the circumstances, as illustrated by Springboard and Shrimp. Of itself, the map embodies the realisation that 'there is no single set or sequence of steps followed in all investigations' (Lederman, et al., 2014, p. 68).

Understanding of the ideas in the map is demonstrated in practice during trials or iterative working; the 'thinking behind the doing' becomes evident. As noted earlier, trials are conducted prior to main data collection and are important to determine the range and interval of the IV, the means of control for confounding variables, and how best to deal with any variation in the data to see if a pattern can be distinguished. The investigator tends to work more iteratively at the start of an investigation and then more linearly once the quality of the data has been established, although decisions about anomalous data, the size of the sample or number of repeats cannot be pre-determined and must be considered, in relation to the data collected and whether it is good enough for the claim, throughout the whole investigation. Investigation, as Hodson (2014, p. 13) states 'is an organic, dynamic, interactive activity, a constant interplay of thought and action'. This explains how different investigators, even if approaching a problem in a similar way, may make different justified decisions in response to the ideas – reflecting their 'constraining assumptions' (Fortus, 2009, p. 86) - with concomitant effects on their data.

The inter-play of these ideas in relation to context is illustrated by different approaches to repeating readings (Heinicke & Heering, 2013). In Springboard the reliability of the data, requires sufficient repeated measurements to 'capture' the variation and give a small enough standard error. However, in circumstances where the event is highly repeatable and specialised instruments have a very small degree of uncertainty, repeating until the same measured value occurs consistently (say, twice) is a sensible approach (e.g. many titrations in chemistry). Any small differences between repeated readings can be ascribed to 'operator error' and in such circumstances practicing the technique leads to consistency. The latter approach corresponds to what Buffler, Allie, Lubben and Campbell (2001) term as 'point reasoning', but we must be careful about classifying this as a general misconception about repeated readings – all depends on the context.

Of course, the formal write up of an investigation may not reflect the iterative working and may just give a linear account, reporting the conditions under which the data were collected without the background story of the preparatory trialling. In this

sense the conventions of formal write ups, although an efficient means of communication, taken at face value, misrepresent scientific practice.

Substantive knowledge and scientific practice are inseparable

The map emphasises the intimate integration of substantive knowledge (the shadowed concept boxes) with scientific practice. Neither stands alone, each is only as good as the other. The production of data is conceived within, is guided by and uses instruments that depend on existing substantive knowledge. The soundness of substantive knowledge depends on the quality of the originating data as evidence. We have already touched on the relationship between theory and evidence. Student engagement in interpretation, evaluation and argumentation has been strongly advocated in recent research literature (Kuhn, 1993; Lubben et al., 2010). However, the argumentation research tends to focus on the fit between theory and evidence without examining the quality of data behind the 'evidence' (Gott & Duggan, 2007). Although this allows students to learn about argumentation *per se* it provides little opportunity for them to engage with the ideas in the map. For example, in a socio-scientific context about the funding of a zoo (Osborne, Erduran & Simon, 2004, p. 1009) the information furnished for the argument is that:

> … 'some animals wouldn't be able to breed in the wild' and there is a warrant supplied that this is because 'they may not have enough food.' This claim is further supported or elaborated by the claim that 'the animals need a safe place to live' and the data to support this claim are that otherwise 'they will be at risk from predators'.

All of the claims are accepted without evidence to support these assertions. Even when empirical data are provided, activities that focus on students' reasoning and their use of evidence (rather than opinion or inference) to support explanations may not provide opportunities for the quality of the data to be questioned. For instance, in detailed work by Berland and Reiser (2009) students' arguments about population numbers were deemed to be more persuasive when they included statements like 'research showed', suggesting some acceptance of the authority of the research and the quality of the data presented seemed not to be questioned.

Duschl and Osborne (2002, p. 55) note that 'the challenge is to provide teachers and students with tools that help them build on nascent forms of student argumentation to

develop more sophisticated forms of scientific discourse'. We would suggest that the ideas about the quality of data are indispensable and that the whole of our map is addressed across argumentation activities. These tools move towards giving students the means to evaluate 'the *goodness*, the *normative status*, or *epistemic forcefulness* of candidate reasoning for belief, judgement, and action' (Siegel, 1995, p. 162, emphasis in original) and are therefore important in both the construction and critique of claims (Ford & Foreman, 2006).

Overall, the appreciation of ideas about the quality of data and the relationship between substantive knowledge and scientific practice lie at the heart of understanding about the nature of science (NOS). The specification of what might be appropriate at school level is under debate (see for instance Abd-El-Khalick, 2012; Hodson & Wong, 2014). There is no space to enter into this debate here but we suggest that an understanding of our map would go a long way towards a desirable understanding of the NOS; a contention seemingly supported by Kind's (2013b) analysis of NEAP, PISA and TIMSS science assessment scales. Indeed, as well as informing curriculum developers and awarding bodies, we believe the map would serve as a useful instrument for teachers when developing students' epistemic understanding about scientific knowledge. Here, reference to the history of science can help to illustrate the interaction of scientific knowledge with scientific practice. For example, early work on the volumes of gases did not identify temperature and pressure as confounding variables (Conant, 1957). A current example where we are as uncertain about the phenomena as our predecessors were is whether data are good enough evidence for the existence of gravity waves or better explained by space dust (Cowen, 2014).

Implications for the role of practical work in science education

Viewing scientific practice as a network of ideas to be understood has significant implications for the role of practical work in science education, its specification in curricula and its assessment (Roberts & Gott, 2006). By practical work we mean activities where students engage with materials and apparatus and make qualitative observations and/or quantitative measurements (after Abrahams and Millar's (2008) definition). There will still be the need to develop manual skills and awareness of

techniques, and for the customary illustrative-type practicals that teachers are familiar with. The purpose of illustrative practicals is to acquaint students with the variables science uses to describe the world and so support the learning of substantive knowledge and theory. Students need to learn how to recognise objects (for example, plant and animal cells when viewing suitably prepared samples through a microscope) or to recognise phenomena (such as dissolving or boiling). Illustrative practicals also exemplify known relationships between variables. To ensure the intended outcome, these can be presented to students as a 'recipe' wherein all the decisions about the design to establish a valid pattern have already been made. In the physical sciences, matters can be contrived to give a small variation in comparison to the effect of changing the IV so that only a few repeated measurements of the DV suffices (three, at most, will do). In biological contexts, where variation is inherent, collection of large datasets - often by many students pooling their results - is often needed to illustrate a relationship unambiguously.

Of course, in a recipe-type practical teachers can make a point of analysing the reasoning behind the design, but the first priority is developing the substantive understanding and one must be mindful of overloading students. In contrast to illustrative practicals there is also the need for activities aimed at developing students' understanding of the quality of data; i.e. the ideas in our map. This will involve students carrying out their own scientific investigations so that they can make their own decisions, applying their understanding of the ideas on the map. In the first instance, we would suggest that teachers choose contexts which do not make high demands on specialised substantive understanding and where the outcome is not part of prescribed disciplinary content, better still if the outcome is genuinely unknown. This allows the focus to be on getting good enough data to tell us something and not agreement with a 'right answer' (Allen, 2011) – a practice that can give a false impression of science (Driver, Newton & Osborne, 2000). Springboard is a good example, which can be extended by adopting it as a measuring instrument to address issues of calibration. Freed from substantive content imperatives, a wide range of imaginative contexts with appeal to students are possible: for example, investigating the relationship between the spins of an egg and the length of boiling time (see Gott, Foulds, Roberts, Jones & Johnson, 1999). Demonstration of an understanding of the interaction amongst the ideas in the map is seen when students carry out trials and

work iteratively in response to the data – making nuanced decisions as they work – which are not features common to illustrative practicals. Shrimp provides opportunities for many different decisions not encountered in most lab-based contexts, thus providing an opportunity for students to develop an understanding of evidence in 'naughty world' contexts where data is 'messy' (Lambert & Reiss, 2014).

Since ideas about the quality of data have not been explicitly and systematically addressed in customary school science (Abrahams & Reiss, 2012), we have little empirically-derived data on how students take to them and how a progression in scientific investigation might look. However, we can identify features of an investigation which, from our experience, affects the demand in applying the ideas. A continuous IV entails more decision making than a categoric IV. A reliable DV is easier to deal with than a wayward one. Large changes in the DV and especially in conjunction with high reliability make for a less demanding decision on the number of repeats (i.e. again, the typical 3 may well do). Situations where confounding variables need to be matched are more challenging than those where they can be manipulated. Sampling of objects with significant variation within their kind brings an additional consideration. How such features of an investigation combine to give an overall difficulty is an empirical question. Even if non-specialised, the substantive context will also be a major factor and it seems likely that a graduated bank of investigations will need to be built up through a case law approach. Of course, in practice, the overall difficulty of an investigation activity can be adjusted by how many decisions are left open to the students. Since the focus is on learning ideas, non-practical teaching activities will also be appropriate; in our experience, explicit teaching of ideas from distinct sections of the map, with ample opportunities for students to discuss the effects of potential decisions in relation to real data and the quality of claims (their own or others') is valuable.

Once students comprehend ideas about the quality of data, as their substantive knowledge develops they should be able to conduct profitable investigations in increasingly more specialised scientific contexts, thereby consolidating their substantive and quality of data understandings in conjunction. This could be seen to be moving towards the goals of inquiry-based learning (Minner, Levy & Century, 2010) and indeed with good enough data students should be able to find out certain relationships between variables for themselves. However, we must emphasise that

investigative success is contingent upon already established understanding about the quality of data and sufficient substantive understanding to allow its realisation in context (Glaesser, Gott, Roberts & Cooper, 2009). We concur with Hodson (2014) that inquiry-based learning cannot be the principal method for teaching substantive knowledge. Carefully thought through illustrative practical work is needed to support the development of important substantive ideas. Here we must acknowledge that research casts doubt upon the value of much practical work in developing substantive understanding (Abrahams & Millar, 2008; Abrahams & Reiss, 2012). However, it seems logical that first-hand experiences of phenomena ought to support learning and we suggest their apparent ineffectiveness calls for a rethink of many traditional activities, informed by the research into students' substantive understanding (Johnson & Tymms, 2011). It is also important to appreciate that many students will not 'pick up' ideas relating to the quality of data simply by following numerous 'recipes' (Gott & Duggan, 1996). The ideas need direct teaching with a combination of illustrative activities (practical and non-practical, as with substantive ideas), and whole and part investigations where students consider the interplay between the ideas in decision making (see, for instance, Campbell, 2010; Roberts, 2004). Once students have developed this basic understanding in different contexts they will, arguably, be in a better position to understand the diverse practices and conventions employed across the sciences (Hodson & Wong, 2014).

In short, teachers must be very clear about the purposes of any particular practical activity (Abrahams & Millar, 2008), which should be more than a way of appeasing students.

Conclusion

The focus on understanding evidence in recent curriculum documents represents an ontological shift in the characterisation of scientific practice; it is not compatible with a process view with its focus on 'doing' certain skills and processes, and therein lies the danger to its realisation. If the teaching of scientific practice continues to be viewed as associated with processes, the conceptual basis for understanding evidence - the core ideas and their relationships which are shown on the concept map – will be misrepresented and will be lost in curriculum implementation. The ideas of evidence will not be at the explicit heart of teaching and assessment of scientific practice.

Deep-rooted customs in school science practice, with its association with a single scientific method, may result in readers of curriculum documents interpreting them in old ways. We would venture to warn that without the ontological shift exemplified here, little may change in practice despite everyone's best intentions.

Our map has gone through many iterations and we make no claims about the version presented here. Its main purpose is in its being a concept map and so emphasising that scientific practice is about ideas to be understood and that can be specified – just like substantive knowledge – and that its teaching should follow accordingly so as to develop learners' understanding of the relationships expressed in the map.

The map emphasises that science is a single entity where substantive knowledge and scientific practice are different facets when viewed from different perspectives. While such facets are useful in describing science and specifying frameworks (e.g. DfE, 2014; NRC, 2012) we must be wary about making divisions in the body of science where none exist. The reification of such a division in the past, especially between substantive as 'conceptual' and practice as 'process' has perhaps not been helpful for science education.

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H-L. (2004). Inquiry in science education: International perspectives. Science Education, 88(3), 397–419.

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education, 30*(14), 1945-1969.

Abrahams, I., & Reiss, M. J. (2012). Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching, 49*(8), 1035-1055.

Achieve, Inc. (2013). *Next generation science standards.* Retrieved from http://www.nextgenscience.org/next-generation-science-standards

Allen, M. (2011). Theory-led confirmation bias and experimental persona.

Research in Science and Technological Education, 29(1), 107-127.

Australian Curriculum, Assessment and Reporting Authority. (2012, January). *The Australian Curriculum: Science.* Retrieved from http://www.australiancurriculum.edu.au/Australian%20Curriculum.pdf?Type=0&s=S&e=ScopeAndSequence

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55.

Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2001). The development of physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education, 23*(11), 1137–1156.

Bybee, R. (1997). *Achieving scientific literacy: From purposes to practical action.* Portsmouth: Heinemann.

Campbell, P. (Ed.). (2010). *The language of measurement: Terminology used in school science investigations.* Hatfield: Association for Science Education (on behalf of ASE-Nuffield).

Cañas, A.J., Novak, J.D. & Reiska, P. (2015) How good is my concept map? Am I a good cmapper? *Knowledge Management & E-Learning*, 7(1): 6–19. Cavagnetto, A.

R. (2010). Argumentation to foster scientific literacy: A review of argument interventions in K-12 science contexts. *Review of Educational Research, 80*(3), 336-371.

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science, 69*(3), 447-451.

Conant, J. B. (1957). The overthrow of the phlogiston theory: The chemical revolution of 1775-1789. In J. B. Conant & L. K.Nash (Eds.). *Harvard case histories in experimental science*, volume 1. Harvard: Harvard University Press.

Cowen, R. (2014). Doubt grows about gravitational waves detection. *Scientific American.* Retrieved from http://www.scientificamerican.com/article/doubt-grows-about-gravitational-waves-detection/

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582–601.

DeBoer, G. E. (2011). The globalisation of science education. *Journal of Research in Science Teaching, 48*(6), 567–591.

Department for Education. (2014) *National Curriculum in England: science programmes of study.* Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*(1), 39-72.

Fairbrother, R., & Hackling, M. (1997). Is this the right answer? *International Journal of Science Education, 19*(8), 887-894.

Fletcher, J. (2014). *Spurious correlations: Margarine linked to divorce?* Retrieved from http://www.bbc.co.uk/news/magazine-27537142

Ford, M. J., & Foreman, E. A. (2006). Redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30.* 1-32.

Fortus, D. (2009). The importance of learning to make assumptions. *Science Education, 93*(1), 86 – 108.

Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009). The roles of substantive and procedural understanding in open-ended science investigations: Using fuzzy set Qualitative Comparative Analysis to compare two different tasks. *Research in Science Education, 39*(4), 595-624.

Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 791-806.

Gott, R., & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in Science & Technological Education, 25*(3), 271–91.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (n.d.). *Research into understanding scientific evidence.* Retrieved from http://community.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Gott, R., Foulds, K., Roberts, R., Jones, M., & Johnson, P. (1999). *Science Investigations: 3.* London: Collins Educational.

Gray, R. (2014). The distinction between experimental and historical sciences as a framework for improving classroom inquiry. *Science Education, 98*(2), 327-341.

Haigh, M., France, B., & Gounder, R. (2012). Compounding confusion? When illustrative practical work falls short of its purpose - A case study. *Research in Science Education, 42*(5), 967–984.

Hall, B. M. (2010). *Teaching uncertainty: the case of climate change.* Unpublished PhD Thesis, University of Gloucestershire, UK.

Heinicke, S., & Heering, P. (2013). Discovering randomness, recovering expertise: the different approaches to the quality in measurement of Coulomb and Gauss and of today's students. *Science and Education, 22*(3), 483–503.

Hodson, D. (2014). Learning science, learning about science, doing science: Different goals demand different learning methods. *International Journal of Science Education,* DOI: 10.1080/09500693.2014.899722

Hodson, D., & Wong, S. L. (2014). From the Horse's Mouth: Why scientists' views are crucial to nature of science understanding. *International Journal of Science Education,* DOI: 10.1080/09500693.2014.927936

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28 – 54.

Jenkins, E. (2009). Reforming school science education: A commentary on selected reports and policy documents. *Studies in Science Education, 45*(1), 65–92.

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "Doing science": Argument in high school genetics. *Science Education, 84*(6), 757–792.

Johnson, P., & Papageorgiou, G. (2010). Rethinking the introduction of particle theory: A substance-based framework. *Journal of Research in Science Teaching, 47*(2), 130-150.

Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry relating to the concept of a substance. *Journal of Research in Science Teaching, 48* (8), 849-984.

Jones, L. R., Wheeler, G., & Centurino, V. A. S. (2013). TIMSS 2015 Science Framework. In I.V.S. Mullis, & M.O. Martin, (Eds.). *TIMSS 2015 Assessment Frameworks* (Chp 2). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/frameworks.html

Kinchin, I. M. (2010). Solving Cordelia's Dilemma: Threshold concepts within a punctuated model of learning. *Journal of Biological Education, 44*(2), 53-57.

Kinchin, I. M. (2015). *Prof. Kinchin's Blog: Musings on Academic Development.* Re: Excellence and elegance in concept mapping. Retrieved from https://profkinchinblog.wordpress.com/

Kinchin, I. M. & Hay, D. B. (2007). The myth of the research-led teacher.

Teachers and Teaching: theory and practice, *13*(1), 43-61.

Kind, P. M. (2013a). Conceptualising the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education, 97*(5), 671–694.

Kind, P. M. (2013b). Establishing Assessment Scales Using a Novel Disciplinary Rationale for Scientific Reasoning. *Journal of Research in Science Teaching, 50*(5), 530–560.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning of scientific thinking. *Science Education, 77*(3), 319–337.

Lambert, D., & Reiss, M. J. (2014). *The place of fieldwork in geography and science qualifications.* London: Institute of Education.

Laugksch, R. (2000). Scientific Literacy: A conceptual overview. *Science Education, 84*(1), 71-94.

Lederman, J. S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A., & Schwartz, R. S. (2014). Meaningful assessment of learners' understandings about Scientific Inquiry - The Views About Scientific Inquiry (VASI) questionnaire. *Journal of Research in Science Teaching, 51*(1), 65–83.

Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting development of the epistemology of inquiry. *Cognitive Development, 23*(4), 512–529.

Lubben, F., Sadeck, M., Scholtz, Z., & Braund, M. (2010). Gauging students' untutored ability in argumentation about experimental data: A South African case study. *International Journal of Science Education, 32*(16), 2143–2166.

Martin, I. (2011). Literacy as metaphor and perspective in science. In C. Linder, L. Östman, D. A. Roberts, P-O. Wickman, G. Erickson, & A. MacKinnon (Eds.), *Exploring the landscapes of scientific literacy* (pp. 90-105). Abingdon, Oxon: Routledge.

Millar, R., & Driver, R. (1987). Beyond Processes. *Studies in Science Education, 14*(1), 33-62.

Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance. *Research Papers in Education, 9*(1), 207–248.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future. A report with ten recommendations.* London: King's College London.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction - What is it and does it matter? Results from a research synthesis Years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496.

National Research Council. (2007). *Taking science to school: Learning and teaching science in Grades K-8.* Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Novak, J. D. (2010). *Learning, creating and using knowledge*.(2nd edn). Oxford: Routledge.

Novak, J. D., & Cañas, A. J. (2007). Theoretical origins of concept maps, how to construct them and use them in education. *Reflecting Education*, 3(1), 29-42.

Novak, J., & Gowin, D. (1984). *Learning how to learn*. Cambridge: Cambridge University Press.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015: Draft science framework.* Paris: author.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Roberts, D. A. (1982). Developing the concept of 'curriculum emphases' in science education. *Science Education, 66*(2), 243–260.

Roberts, D. A. (2007). Scientific literacy/science literacy. In S.K. Abell, & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 729–780). Mahwah, NJ: Lawrence Erlbaum

Roberts, D. A. (2011). Competing visions of scientific literacy: the influence of science curriculum policy image. In C. Linder, L. Östman, D. A. Roberts, P-O. Wickman, G. Erickson, & A. MacKinnon (Eds.), *Exploring the landscapes of scientific literacy* (pp. 11-27). Abingdon, Oxon: Routledge.

Roberts, R. (2004). Using different types of practical within a problem-solving model of science. *School Science Review*, 85(312), 113-119.

Roberts, R., & Gott, R. (2003). Assessment of biology investigations. *Journal of Biological Education*, 37(3), 114-121.

Roberts, R., & Gott, R. (2006). Assessment of performance in practical science and pupil attributes. *Assessment in Education*, 13(1), 45-67.

Roberts, R., & Gott, R. (2010). Questioning the evidence for a claim in a socio-scientific issue: An aspect of scientific literacy. *Research in Science and Technological Education, 28*(3), 203–226.

Royal Society (n.d.). *The Royal Society: History*. Retrieved from https://royalsociety.org/about-us/history/

Schalk, H. H., van der Schee, J. A., & Boersma, K. T. (2013). The development of understanding of evidence in pre-university biology education in the Netherlands. *Research in Science Education, 43*(2), 551–578.

Shymansky, J. A., Yore, L. D., Treagust, D. F., Thiele, R. B., Harrison, A., Waldrip, B. G., Stocklmayer, S., M., & Venville, G. (1997). Examining the construction process: A study of changes made in Level 10 students' understanding of classical mechanics. *Journal of Research in Science Teaching, 34*(6), 571-593.

Siegel, H. (1995). Why should educators care about argumentation? *Informal Logic, 17*(2), 159-176.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology and Evolution, 24*(9), 467-471.

Tytler, R. (2007). *Re-imagining science education: Engaging students in science for Australia's future.* Victoria: Australian Council for Educational Research.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education, 34*, 254-284.

Figure legends / list of captions

Figure 1. A concept map with the focus question "What is the 'thinking behind the doing' for determining the validity of data?" (Concepts directly informed by substantive knowledge are highlighted with a shadow on the box.)

Figure 2. The set up for Springboard. When the string is cut, the toy figure is launched upwards.

Figure 3. A freshwater stream survey typical of Shrimp. With kind permission of the Field Studies Council.

Footnotes

[i] Other terms are in common usage; for example variables can be referred to as factors; the IV and DV being input and output factors. Some descriptions such as 'the thing you measure' for the DV can be misleading since all variables' values are measured.

[ii] This involves scuffling at a regular intensity for a known period of time in a defined area to dislodge organisms from the substrate for collection in a net immediately downstream. Some people may consider such a technique to be so disruptive that they would not use it – an interesting ethical dimension that should be considered.

[iii] Proxy measures are very important in 'historical' sciences, such as geology/earth science and in the study of climate change e.g. tree rings and ice cores as proxy measures of climate conditions (see Hall, 2010).

[iv] Identification error is a potential threat to the quality of 'citizen science' surveys and checks on the data are often built into the procedure (Silvertown, 2009).

[v] Non-parametric tests of difference are important for data that are not normally distributed.

[vi] Not only by 'experiment' – this refers to science's empirical basis.