The experimenter expectancy effect: an inevitable component of school science?

ABSTRACT: A medium-scale quantitative study (n=99) found that 10-11 year-old pupils dealt with theory and evidence in notably different ways, depending on how the same science practical task was delivered. Under the auspices of a part-randomised and part-quasi experimental design, pupils were asked to complete a brief, apparently simple task involving scientific measurement. One half of the sample carried out the task in a naturalistic whole class context; the other half worked as lone experimenters in solitary conditions where accuracy of measurement was promoted. In the whole class setting pupils exposed to an illustrative lesson displayed behaviour indicative of experimenter expectancy, tending to differentiate theory and evidence to a lesser degree than pupils who experienced an enquiry lesson. In addition, during the illustrative lesson many of the pupils were biased towards their theories in ways that lay well beyond those intended by the research design. In the solitary setting pupils performed equally well with both illustrative and enquiry treatments. Implications are discussed in the light of the problems of excessive pupil theory/data-ladeness. Finally, the advantages and disadvantages of exposing young learners to more authentic versions of professional science are considered.

Introduction

The process of scientific thinking involves the interplay of two elements that have been historically regarded as separate entities – theory and evidence (Kuhn & Pearsall, 2000). The modern view holds that both are vital, helping to give science a robustness that is lacking in other ways of thinking, such as reflected in religious assertions where empirical evidence is deemed unnecessary. 'Theory' represents the product of internal human mental activities; 'evidence' relates to events taking place in the outside sphere that can be sampled via human perception. Descartes famously characterised this duality, emphasising the disconnectedness of the two entities (Alsop, 2005).

Although debate continues about what precisely constitutes the scientific method (Lawson, 2010), all varieties of modern science are informed by a post-positivist philosophy that assumes an evidence-based position. But throughout science's evolution physical evidence has been given varying degrees of importance. The medieval scholar Thomas Aquinas held that since science facts and laws were part of a Nature created by God, once discovered they can never be challenged by new evidence (Chesterton, 1933); he was excessively *theory-led*. Later, in the 17th century Francis Bacon recognised that much of medieval science was static, with laws being written in stone, and preconceived notions led to biased conclusions. He advocated instead an inductive method which required an objective approach: expectations regarding the outcome of an experiment were forbidden, with theories being

derived directly from observations of physical phenomena (Priest, 2007). To Bacon, prior knowledge was not considered relevant and evidence was sacrosanct – he was excessively *data-led*. Directed by thinkers such as Newton and Galileo this developed into the contemporary view of a hypothetico-deductive scientific method, where hypotheses are formulated from a theory then tested during experimentation, either being accepted or rejected as a consequence (Poletiek, 2001). These later scientists were neither excessively theory nor data-led.

For science to work both theory and evidence need to be in agreement - if they diverge, then one of them is incorrect and must be discarded. As stated, a theory can be rejected when using a hypothetico-deductive method as a result of convincing experimental evidence that refutes that theory, and an alternative then proposed. On the other hand, when theory and evidence do not correspond it is sometimes due to poor methodology yielding invalid data, or unknown confounding variables, and once this is recognised then the evidence is rejected and the theory remains viable. Havdala and Ashkenazi (2007) call this *the informed view*.

As happened with Aquinas and Bacon, modern scientists aligned with a hypothetico-deductive approach have in a similar way sometimes given inappropriate weighting to either theory or evidence (Dunbar, 2000; Greenwald, Pratkanis, Lieppe & Baumgardner, 1986). If a theory is strongly believed prior to experimentation, then confirmation bias can ensue where evidence that refutes that theory is sidelined, and only data that support the theory are recorded, resulting in a self-fulfilling prophesy where the theory survives. This is taking an overly theory-laden approach. Conversely, excess import can be laid upon experimental evidence, with scientists having an overly data-laden mind set. British Empiricists such as Bacon awarded undue importance to their data which they viewed as being objectively pure and so irrefutable (Priest, 2007), mechanically disproving any opposing theory without giving thought to experimental error or unconscious human biases.

Much of school science involves illustrative, or verification practical work where the class firstly learns the scientific concepts behind an experiment then carries out a laboratory activity in order to demonstrate that principle in action (Nott & Smith, 1995; Rogan & Aldous, 2005). During these lessons pupils are generally aware of what the outcome of the activity will be, because it is based on known theory, their task being simply to confirm that theory. If they fail to do this then their data must be wrong, and teachers usually cite learners' experimental failure as the reason. By design these activities only serve to blur the boundary between theory and evidence in pupils' minds, but are thought to be necessary in order that correct science is learned (Nott & Smith, 1995). Conversely, with more open-ended, enquiry activities typically pupils are unaware of the outcome. They make predictions or hypotheses then test them experimentally, as would a professional scientist, and the usefulness of these more authentic portrayals of science in the classroom have been lauded (e.g.

Blanchard, Southerland, Osborne, Sampson, Annetta & Granger, 2010; Dean & Kuhn, 2007; Fairbrother & Hackling, 1997; Marx, Blumenfeld, Krajcik, Fishman, Soloway, Geier *et al*., 2004).

As will be appreciated from the preceding discussion there are differences between the ways professional scientists and school pupils conduct their science, although both groups follow procedural rules that require any conclusions to be evidence-based. This assumption still holds during illustrative lessons in school where the purpose is to confirm a known theory. However, pupils have demonstrated similarly errant attitudes to their professional counterparts towards theories and the evidence they collect during class experiments. Even during enquiry experiments, when faced with a theory/evidence disparity they tend to assume an overly theory-laden mind set and automatically reject their empirical results out of hand (e.g. author, 2010; Austin, Holding, Bell & Daniels, 1991; Gunstone & Champagne, 1990; Lubben & Millar, 1996; Zimmerman, Raghavan & Sartoris, 2003). Pupils may have difficulties switching between the positivistic epistemology of the more common illustrative practicals and the constructivist epistemology of the usually rarer hypothetico-deductive open ended enquiry tasks (author, 2011). This might create a tendency amongst pupils towards being positivistic and overly theory-laden in all of their scientific enterprises, even during enquiry activities when there is no authoritative theory to confirm. An extreme position involves young experimenters believing that the evidence they have been instructed to collect is merely another version of the theory under investigation (Foulds, Gott & Feasey, 1992; Havdala & Ashkenazi, 2007; Kuhn, Amsel & O'Loughlan, 1988; Zimmerman, 2007). These pupils assume it is their role to actively seek out results that confirm a single, predetermined outcome, and if they collect results that refute that outcome, they have failed the task. They have not properly differentiated theory and evidence; instead, they believe that they represent the same entity.

Dialogue between peers is generally considered to be fruitful in science lessons, particularly during experimentation where pupils typically work in small groups (e.g. Mercer, Dawes & Staarman, 2009). However, within the literature there is some support to the view that the influence of peers can exacerbate theory/evidence non-differentiation, making theory-led behaviours more or less inevitable.

"*In many carefully organised experiments the discovery is made by the quickest member of the class, often the noisiest, who then provides the rest with the answer*" (Wellington, 1981, p168).

These social factors, much studied in the psychology genre, can manifest as a band wagon effect especially when data are ambiguous (e.g. Asch, 1951; Baron, Vandello & Brunsman, 1996; Stangor, 2004; Wood, 2000). Pupils sometimes have so little regard for the empirical evidence they have personally collected during a science lesson that they habitually copy the results of collaborators in the class, if they think that they have the 'right answers' (Atkinson, 1990; Del Carlo & Bodner,

2004; Rigano & Richie, 1995). Alternatively, peers can compete with each other in a chase for the right answer which triggers theory-led behaviours such as inventing results that correspond with their theory, or manipulating apparatus to give the desired outcome (author, 2011).

Theoretical background

Rosenthal performed a series of classic psychology experiments in the 1960s that investigated how the prior expectations of experimenters influenced their subsequent actions during data collection and inference making (Rosenthal, 1966). For one study, he accrued two groups of adult volunteers and asked them to perform an experiment that involved timing lab rats as they negotiated a maze. One group was told to expect that their rats were able to learn a maze quickly (maze-bright rats), and the other group were told theirs would not learn the maze easily (maze-dull rats). Consequently, the maze-bright group of participants recorded significantly faster times for their rats than the maze-dull group (mean $p < 0.01$). However, the rats used for both groups belonged to the same population and so there should have been no differences.

Rosenthal's work demonstrated the powerful influence of scientific expectations that have been inadequately bracketed during data collection; he called this the *experimenter expectancy effect*.[i] Rosenthal concluded that the effects in his study were unintentional and experimenters were unconsciously handling the rats so to increase their speed around the maze. Participants had been overly theory-laden and had without realising it failed to differentiate adequately between theory and evidence, even though they themselves might have thought they were experimenting fairly. Other than the field of animal behaviour, the experimenter expectancy effect has also been demonstrated in other settings, including the legal (e.g. Martindale, 2005) and the medical professions (e.g. Wigal *et al*., 1997) as well as in more everyday contexts (e.g. Gilovich, 1991).

The science education literature contains some references to how experimenters can be biased by their own expectations, though this has always been a neglected area of research (e.g. author, 2007; Chinn & Brewer; 1998; Hainsworth, 1956 & 1958; McCormas & Moore, 2001; Rigano & Richie, 1995; Watson, Swain & McRobbie; 2004; Zimmerman *et al*., 2003). The authors of one study of experimenter expectancy in school biology pupils (aged 14-19) describe the effect potentially being a universal problem that can apply to almost any kind of scientific measurement in the classroom (McCormas & Moore, 2001). The current study replicates the principle behind Rosenthal's rat research, with two groups of school pupils being allowed to have different expectations of events that should in fact yield similar results, in an attempt to investigate what happens if theory and evidence fail to correlate. The findings will be discussed with reference to the theory and data-ladeness of participants. In this respect, the study represents work that is original to science education.

There has been recent debate in the literature regarding the efficacy or otherwise of enquiry lessons over more didactic forms of teaching science. Blanchard *et al*. (2010), reviewing a wide range of previous research concluded that any reported advantages of an enquiry approach have been at best mixed, or inconclusive. Interestingly, this equivocal research base is juxtaposed alongside the major emphasis, involving huge resources of manpower and funding, placed on enquiry-based science over the past 20 years within school curricula in the UK, US, Australia, and elsewhere. The current study did not examine any pedagogical superiority/inferiority of enquiry lessons over illustrative lessons with respect to substantive learning, as it did not attempt to measure pupils' constructions of science concepts. Instead, the focus was on pupils' procedural behaviour, specifically their treatment of theory and evidence, and as such intends to contribute to the continuing debate over the worth of enquiry in school science.

Methodology

Preamble

The problems associated with pupils' failure to differentiate theory and evidence act as a barrier to learning science, particularly the proper procedural rules of experimentation. The current research aimed to compare pupils' behaviours during different types of practical work in order to study how well they were able to differentiate between theory and evidence. This was operationalised in the form of two randomised educational experiments, where the effects of two different treatments were compared within the bounds of each experiment.

In the first experiment pupils experienced either one of two treatments when being taught as a whole class. The first treatment was illustrative in type where pupils were asked to undertake a practical task, but were told one of the outcomes at the outset (theory-led group). The second treatment took the form of an enquiry exercise where pupils carried out the same practical activity as used with the first treatment, though were not told any outcome (hypothetico-deductive group).

The second experiment involved pupils collecting their data in a side room away from the rest of the class, acting as solo experimenters. As with the first experiment, the first treatment was illustrative (theory-led group) and the second treatment was enquiry in type (hypothetico-deductive group). When compared with the first experiment, the conditions of the second experiment more resembled how professional scientists might collect their data.

Research questions

1. Do pupils treat theory and evidence differently when undertaking an illustrative science activity, compared with an enquiry science activity?
2. Do pupils treat theory and evidence differently when in a whole class context, compared with when they work independently away from their peers?

Participants and sampling

The study was carried out in a mixed gender, non-selective state maintained middle school in the southeast of England. The samples were accrued from year 6 pupils (ages 10-11 years). As shown in figure 1, there were two distinct phases of the research: the whole class phase (49 pupils), and the lone experimenter phase (50 pupils). Within each phase pupils were randomly divided into groups, the theory-led group (T-L), and the hypothetico-deductive group (H-D). In the lone experimenter phase a random error group (R-E) was also formed via the randomisation process. Randomisation was achieved with respect to pupil ability, measured using non-verbal reasoning tests (NFER-Nelson, with standardised scores). The two phases took place in the same school and used pupils of the same age, but drew their respective samples from different cohorts of pupils. As such, any comparisons *between phases* are quasi-experimental in type, and so associations must be made more tentatively than those *within phases*, which accrued randomised samples (see figure 1). After sampling was complete there were found to be no significant ability or gender differences between any of the samples or sub-samples in the study.

Method

a) Whole class phase

Both groups carried out the experiment on the same day over two successive lessons and did not meet between the lessons, so preventing inter-group contamination. The researcher acted as the teacher in both sessions, and each lesson took around 20 minutes to complete.

Theory-led group. Pupils were asked to carry out a simple exercise that involved measuring and recording the temperatures of three different plastic cups of water using a glass thermometer that had an analogue scale. The cups themselves were identical physically and contained the same volume of water (one third full). Each of the three cups was lagged with a different material, namely fur, felt, or aluminium cooking foil, and had no lids. The water had been left to equilibrate overnight in the room

and so was at ambient temperature during the study.[ii] The cups were explicitly labelled in a large font with the name of the lagged material.

The lesson began with pupils pairing up then completing a short worksheet question that required them to make predictions, with free discussion being encouraged amongst the class (figure 2). The item was presented in the form of a bluff question that erroneously implied materials at ambient temperature would actively warm a cup of 'cold' water.[iii] In order to inform their judgements during the prediction phase, pupils had a chance to handle examples of the materials, not attached to any of the cups. In an attempt to generate experimenter expectancy, after making predictions the class were erroneously informed by the teacher that out of the three choices, fur was actually the best material to actively heat water, and this was the reason why animals in a cold climate had thick fur. Pupils who had made the 'correct' prediction were praised, and to further enhance any confirmation bias pupils were (again erroneously) advised that the water had been in the cups for long enough to allow it to get warm.

Keeping in pairs, pupils then commenced with the practical task. Each pair had the three cups in front of them and had to use a single thermometer to measure the temperature of each. No further advice was given with regards to carrying out the activity. All results were recorded on the worksheet (figure 2), that had the prediction clear in view on the same page. After collecting the three results, pupils were asked to summarise their findings on the same worksheet under 'conclusions' by ranking the three materials in order of effectiveness at heating 'cold' water (1^{st}, 2^{nd}, 3^{rd}). However, immediately before making their conclusions they were asked to write fur at the top of the conclusion list, because that should be the best material at warming the water. Thus, the T-L lesson was semi-illustrative since pupils were given one outcome (fur) and had to determine two unknown outcomes (felt and foil) by experimentation.

Hypothetico-deductive group. After making predictions in an identical manner to the theory-led group, pupils were then asked to measure the three cup temperatures. Unlike with the theory-led pupils, there was no prior discussion of prediction choices by the teacher; particularly, no 'right answer' was publicly or privately told to pupils. As with the theory-led group, pupils were informed, to help develop confirmation bias, that the water had been in the cups for long enough to become warm. Pupils then carried out the experiment in exactly the same manner as the theory-led group; this included the writing of results and conclusions, but pupils were given free choice when ranking their conclusions.

The H-D lesson was not 'true' open-ended enquiry, since unknown outcomes had to be determined in a context controlled by the teacher. Havdala & Ashkenazi (2007) describe a typology of

enquiry where the higher the level the more pupil autonomy is allowed. The H-D lesson in the current research was at level 1 (structured enquiry); true open-ended enquiry tasks lie at level 3.

b) Lone experimenter phase

The lone experimenter phase was carried out in order to examine pupils' behaviours away from any peer influence. Pupils carried out the same practical task as was done during the whole class phase in conditions that more resembled how professional scientists might work. Although scientists operate in collaborative teams, they tend to collect their data alone, making careful measurements and observations free from distractions. This is contrary to school science activities, where pupils tend to work together in groups – data collection is a team effort, often taking place in a noisy, lively environment filled with peer interaction.

Prior to the experiment, pupils underwent the same process as the whole class phase sample, making predictions together with the rest of their group. They were not told any 'right answer' at this stage. One by one, pupils were led into a side room where the experiment was to take place. The apparatus was placed in a well-lit quiet corner of the room away from the researcher's gaze so to create an atmosphere of privacy for experimenters. Unlike with the whole class phase where pupils were issued a single thermometer, to aid accuracy of measurement one thermometer per cup was utilised instead. Accuracy was further promoted by pupils being allowed a brief practice before the actual data collection stage, being shown three flashcards of thermometers at different temperatures. Immediately after each pupil had left, the thermometers were checked to ensure the temperatures had not inadvertently risen due to handling, and none had. The movement of pupils was organised to avoid contamination between those who had completed the experiment and those who had not.

Theory-led group. Each pupil entered the side room alone and their prediction sheet was examined by the teacher. To generate experimenter expectancy, the researcher informed the pupil that the material they had chosen as being most likely to warm the water was actually correct, was praised, and a brief reason was offered, i.e. *this is why foil is used in cooking/polar bears have thick fur/felt is used for slippers*. Pupils were then led over to the apparatus and asked to make observations and record the three cup temperatures on their worksheet; the researcher walked away from the apparatus and left them alone to work. Pupils were not asked to write conclusions as this would had involved writing a rank order of heating effectiveness (in the absence of peers), and since they would have done this immediately after writing the temperatures down, it was thought likely they would simply restate their results.

Hypothetico-deductive group. The procedure was identical to that utilised with the theory-led pupils, except that no 'right answer' was suggested during discussion of the prediction choice.

c) Random error group

In order to study how accurate pupils could be in using thermometers to measure the temperature of water in the absence of any biasing preconception effects, nine pupils were randomly selected during the sampling process of the lone experimenter phase. Similarly, pupils were led one-by-one into a side room then asked to write down the temperatures of 'cold' water in three bare cups *without insulating material*. Pupils were informed that their skill at reading a thermometer was being assessed.

(INSERT FIGURES 1 & 2).

Results

See tables 1-6. ('U/C' represents written entries that were uncodable, which was nearly always because the pupil had not written down any value on the worksheet).

(INSERT TABLES 1-6)

Analysis

The following is a summary of analyses that were applied to the data. The statistical methods utilised were Student's t test, ANOVA, and the chi square test for independent groups; the confidence limits for statistical significance were set at 95%. Note that there were no significant links between either pupil ability or gender with any of the variables discussed hereon in.

a) Whole class phase

Predictions. In both the theory-led (T-L) and hypothetico-deductive (H-D) lessons, pupils' predictions were heavily weighted towards fur as being solely the best material for heating cold water (T-L = 19/22 pupils; H-D = 18/24).[iv] In both groups, fur was chosen more often in preference to foil and felt (T-L $p < 0.0001$; H-D $p < 0.001$).

Results. More pupils in the T-L group recorded fur as being the sole warmest material (T-L 9/20; H-D 4/26 group); this difference between the groups is *not* statistically significant ($p = 0.06$). However, within the T-L group fur was observed as being the sole warmest more frequently than foil or felt ($p < 0.05$), while with the H-D group each material was observed at an equivalent frequency. The mean temperatures that pupils recorded were T-L = 19.1ºC, and H-D = 18.5ºC, which is a significant difference ($p < 0.01$).

Conclusions. More T-L than H-D pupils ($p < 0.01$) chose fur as being the sole warmest (T-L 12/17; H-D 4/20 H-D). As with the results phase, the T-L group had chosen fur as the sole warmest more often than the other materials ($p < 0.0001$), though with the H-D group there were no differences.

Matching predictions with results. There were 8/19 T-L pupils whose prediction ranking matched perfectly with their results compared with 0/26 of H-D pupils ($p < 0.01$).

Matching predictions with conclusions. 5/16 T-L pupils perfectly matched their prediction ranking with their conclusions compared with 1/20 of H-D pupils, which is not significantly different.

Matching results with conclusions. 5/14 T-L pupils recorded conclusions that *did not* reflect their written results; 1/23 in the H-D group did likewise ($p < 0.05$).

Prediction–results–conclusions match. In the T-L group 3 pupils had both observed and concluded precisely what they had predicted; no-one in the H-D group fell into this category (difference not significant).

b) Lone experimenter phase

Predictions. Although metal foil was the commonest material predicted to be the warmest in both groups (T-L = 10/21; H-D = 10/20), statistically, it was not chosen significantly more than fur or felt.

Results. No one material was recorded as being solely the warmest significantly more often than the other two in either the T-L or H-D group - each material was observed at an equivalent frequency. The mean temperatures that pupils recorded were T-L = 23.4ºC, and H-D = 23.3ºC, which is not a significant difference.

Matching predictions with results. In the T-L group, 1/21 had their observations perfectly match their predictions; in the H-D group, it was 3/20. This difference is not significant.

c) Random error group

The mean temperature that random error pupils recorded was 22.9ºC (n=9); this value was significantly lower than the mean temperatures obtained in the lone experimenter phase by both the T-L group (23.4ºC, $p < 0.05$) and the H-D group (23.3ºC, $p < 0.05$). The mean real temperature for the three cups, as measured by the teacher was 22.9ºC.

Temperature ranges

The temperature range for each pupil was their highest recorded temperature minus their lowest recorded temperature and was considered a measure of how accurate the pupil had been when reading the thermometer (table 6). There were no differences between the T-L and H-D groups in the whole class phase; the same held for the lone experimenter phase. When the T-L and H-D groups from each phase were combined, so comparing whether one phase comprised of pupils who were more accurate, there were no differences. The random error pupils had significantly lesser temperature ranges than those of the whole class combined T-L/H-D groups ($p < 0.01$) and also the lone experimenter combined T-L/H-D groups ($p < 0.05$).

Discussion[v]

The study aimed to compare how well pupils would differentiate between theory and evidence under different experimental conditions. Each pupil's theory was an internal construct that represented an external scientific phenomenon, i.e. their current 'best guess' of reality, and was sampled a maximum of twice during the lessons, as their written prediction and conclusion. Their evidence was the temperatures they had recorded (tables 1-4).

a) Whole class phase

Pupils from both T-L and H-D groups predicted fur as being the most likely to heat the 'cold' water the most (86% and 75% respectively, see table 5). Note that prediction-making took place before pupils in the T-L group were erroneously told that fur was the 'right answer'. None of the pupils made the scientifically correct prediction, that all materials would have an equal (zero) effect on changing the temperature of the water.

When it came to data collection, although both the T-L and H-D pupils were observing the same type of scientific phenomena the two groups recorded markedly different results. Although around 50% of pupils in both groups produced results that indicated one of the materials appeared to

be having a superior heating effect when compared with the others, the T-L group tended to record fur as being the most effective (43%). Three of the T-L group had recorded results for foil and felt but had left fur blank. This could be due to having been told by the teacher that fur should be the best heating material, and they considered this to be a foregone conclusion, seeing no need to write a result for fur; none of the H-D group had blank results. With the H-D group, all three materials were reported as the warmest with equivalent frequency. Since the sole independent variable between the two groups was that the T-L pupils had been told to expect a specific occurrence, that fur would have a better warming effect than foil or felt, these data suggest that Rosenthal's experimenter expectancy effect was generated within some of the T-L pupils.

That pupils were influenced by the teacher's explicit stating of a 'textbook' answer is perhaps unsurprising, since this approach would be familiar to them as it is the mainstay of illustrative practical work in school science where the aim is the confirmation of a known outcome. That said, there are problems with theory-led experimentation, discussed previously, and also later in *Implications*. However, the T-L pupils in the current study did not only appear to be influenced by teacher-led expectations. On top of this, it also seems that they allowed their own personal expectations to cloud their data collection. Forty per cent of the T-L group (8 pupils) not only recorded fur as being the warmest material (as they had predicted, and had been confirmed as 'correct' by the teacher), they also perfectly matched all three of their predictions with their three results (e.g. pupil # 1 from Table 1). No-one in the H-D group had done this.

Members of the T-L group were primed by the teacher to expect one result to be the 'right answer' - that fur would come out at the top of the hierarchy. It might be the case that this priming created a general mind-set for confirmation bias to thrive, and even proliferate, which culminated in not only the fur result being unfairly arrived at in order to get the 'right answer', but also other predictions that had not been verified by the teacher as being correct. These pupils were not only generating the outcome that they knew they had to get right, they also did the same with the outcomes that they did not have to get right. These T-L pupils had been more influenced by their theory - what they imagined all the outcomes would be, instead of their evidence - what the outcomes actually were. In a procedural sense, they had been poor scientists. They failed to differentiate between internal theory and external evidence, seeing them both as equivalent entities, with the purpose of the exercise being to produce evidence so that the two corresponded.

When it came to making conclusions, pupils were asked to place the three materials in rank order of effectiveness at heating 'cold' water. Fur was reported as being solely the most effective by 71% of T-L pupils, while no-one chose foil or felt (table 5). The H-D pupils reported the three materials with equivalent frequency. Again, this is perhaps not surprising since the teacher had reminded the T-L pupils just before they made conclusions that fur was the more efficacious material,

although this too demonstrates the experimenter expectancy effect being triggered in response to an overly theory-led approach. When concluding, procedurally, pupils are expected to write statements that fairly reflect the empirical data they have collected. However, a number of them had results and conclusions that did not match, with more of the T-L pupils failing to match (33%, 5 pupils, e.g. pupil # 4), compared with the H-D group (4%, 1 pupil, # 28). Four of the five T-L pupils who had done this opted for the fur-foil-felt hierarchy when writing conclusions, which was the most favoured hierarchy chosen by the T-L class as a whole during the conclusion phase (7/16, 44%). None of these five T-L pupils had recorded fur as being the sole warmest material in their results. It is possible that these pupils were influenced by a general consensus into generating a conclusion that did not fairly reflect their empirical data, because others in the class believed that this hierarchy was the 'right answer'. The bandwagon effect has been noted previously during school science practical work both during data collection and (especially) when making an inference such as an experimental conclusion (e.g. Atkinson, 1990; author, 2011; Rigano & Richie, 1995).

An alternate interpretation is that since none of the five pupils had results that showed fur as being the sole warmest material, they were influenced only by the 'textbook' answer as pronounced by the teacher, since four of them ended up concluding this specific outcome. It might have been a combination effect of both social and 'textbook' pressures. A further possibility involves random carelessness as being the cause for writing mismatched results and conclusions, although this is thought to be unlikely since mismatching hardly occurred in the H-D group (1 pupil, 4%), suggesting that some factor embedded within the T-L lesson was to blame. The current study provided no data with which to confirm or refute these speculations. However, despite the lack of evidence relating to the reasons that lay behind pupils' behaviours, it is proposed that since none of these five pupils were the same as the eight T-L pupils who had previously matched their predictions precisely with results, it appears experimenter expectancy most likely had had a major influence on 13/23 pupils during the T-L lesson. The effect looks to have had a lesser though still notable influence on others in the T-L group who had recorded or concluded fur as being the 'right answer'; for instance 100% of the group wrote fur as being the warmest material in their conclusions (71% as the sole warmest material, 29% as equal warmest).

b) Lone experimenter phase

It is the norm in school science for activities to take place in a lively, interactive, and sometimes distracting environment. Conversely, the experimental conditions for the lone experimenter phase were devised in an attempt to recreate an environment that more resembles how a professional scientist might collect data, using care and accuracy of measurement free from distraction. There were no differences between the predictions or observations of the T-L and the H-D groups that were

statistically significant (pupils did not make conclusions during the lone experimenter phase). In addition, within each group there was no single material that was predicted or measured as being significantly warmer than the others. The T-L and H-D pupils did not differ with respect to any of the variables that were discussed previously with the whole class phase, therefore no expectancy effect was apparent when the two lone experimenter groups were compared (although see random error group results, discussed later).

The current study took the form of two randomised experiments, the whole class and lone experimenter phases. Within each of these phases, independent variables were kept constant apart from the independent variable under study, i.e. a theory-led approach. This enabled valid comparisons to be made within each phase, with any noted differences being likely to be related to the independent variable. However, since *between phases* comparisons are quasi-experimental in type, associations must be made more tentatively than those *within phases*, which were true randomised experiments. That said, data indicate that although the T-L pupils in the lone experimenter phase had performed a similar illustrative task to their whole class T-L counterparts, their results were not as influenced by their knowledge of a 'right answer'. One might have expected the former pupils to have been *more* influenced than the latter. With all the lone experimenters what they had predicted to be the best material was confirmed as being correct by teacher before they collected data, so every personal preconception was reinforced by the textbook answer - instead, the opposite happened.

c) Random error group

Part of the lone experimenter sample (n=9) was used to determine how accurate pupils could be when using thermometers to measure the temperature of water in the absence of any preconceptions due to lagging materials (table 4). Because of the controlled conditions associated with having a single set of apparatus, the temperatures within each cup were known by the teacher during the lone experimenter phase, the random error group included (see table 4). Only one of the nine pupils in the random sample recorded the temperatures with complete accuracy (pupil # 99), which gives an indication of the difficulty pupils of this age group have with this particular skill. However, the pupils' errors did not vary by more than 1ºC, apart from one pupil who registered an error of 1.5 ºC for a single cup. Interestingly, despite these errors, on average the readings obtained by the random group coincided precisely with the average real temperature, as monitored by the teacher throughout data collection (22.9ºC).

In contrast, while some of the pupils within the T-L and H-D groups of the lone experimenter phase read the thermometers with good accuracy, as a whole they recorded higher temperatures than the random group (means were T-L = 23.4ºC, H-D = 23.3ºC, with no significant differences between these two values). The presence of the lagging materials did not make measurement physically more

difficult, so it is likely that there was a degree of experimenter expectancy with both T-L and H-D pupils that culminated in them recording higher temperatures than was actually the case.

In the whole class phase the T-L pupils recorded significantly higher temperatures than the H-D pupils (19.1ºC and 18.5ºC respectively). This may have been an artefact, since the two lessons took place 30 minutes apart and ambient temperatures may have changed slightly; alternately, it might be a further indicator of experimenter expectancy in the T-L group. It was not possible to compare the random group with the whole class phase in this manner. This is because during the whole class phase each group worked in a different part of the room where local temperatures may have varied slightly. Also the researcher noted during trialling that when several thermometers were placed in the same cup of water that they had inherent inaccuracies and showed a range of different temperatures (which is typical of thermometers used in school science). These factors were not confounding since each whole class group used one thermometer, and their three cups were in the same part of the room so any local variations would not matter.

Conclusions

Résumé

In a whole class environment, more pupils who experienced an illustrative lesson (T-L) appeared to be overly attached to theory when compared with pupils who underwent an enquiry lesson (H-D), with many in the former group failing to differentiate theory and evidence. This difference was not apparent when pupils carried out the same task in quiet, more solitary conditions where accuracy of measurement was encouraged free from distraction. This lack of differentiation manifested as the 'textbook right answer' of fur as a superior heating material being recorded and concluded at frequencies greater than chance with the T-L whole class group. In addition, as well as confirming the teacher's answer, 40% of the T-L pupils had also confirmed the rest of their predictions. Different pupils from this 40% had merged theory and evidence when they wrote conclusions that supported the right answer, even when their data refuted that theory.

The current study provided few qualitative data with which to investigate the thinking behind pupils' behaviours. However, the quantitative data give clues, as do the findings of other workers who have investigated similar behaviours, so providing an opportunity to make some tentative comments. In the remainder of this article contemplations are offered concerning the issue of likely pupil motives which help to explain the current research's findings.

Pedagogical effectiveness of the treatments

Can it be argued that the whole class T-L group were poor scientists? On one level they were replicating a procedure that would have been familiar to them, confirming a 'textbook right answer', so were not breaking any procedural rules. Pupils commonly encounter theory/evidence mismatches, and are generally told by the teacher to go with the textbook and put their 'wrong' results down to experimenter error (Claxton, 1986). However, the T-L group were reporting phenomena that should not have actually occurred. When pupils fail to collect the 'right result', they can fabricate a more appropriate substitute (e.g. Keiler & Woolnough, 2002); this is one possible explanation. They may have rejected any anomalous data by simply deciding their experiment was a failure, a justification which has been noted in other studies (author, 2011; Chinn & Brewer, 1998; Gunstone, Mitchell & the Monash Children's Science Group, 1988; Hesse, 1987). Alternately, either consciously or unconsciously, they could have manipulated apparatus or method to ensure the 'right answer' was gained (author, 2011). Possible strategies include warming cups with their hands, and dipping their fingers in the water to test the temperature (one pupil was seen to do this). During science lessons pupils sometimes undertake a 'quick check', where they say '*ok we know what the answer will be but let's do a check anyway* (quickly glances at thermometer), *ok that's right, it's just as we thought'* (author, 2006). These pupils are well aware of the 'right answer' that they should be collecting, and so experience a form of self denial where they suspect that their experiment might be wrong, and so data might refute their theory, and so a careful, longer check is deliberately avoided. On this level, the T-L whole class lesson was inferior pedagogically because it bred experimenter expectancy.

It could equally be argued that the H-D lesson was pedagogically poor since it failed to refute the misconception that materials at ambient temperature can actively warm other ambient materials (first reported by Tiberghien, Sere, Barboux & Chomat, 1983). The random error group's results (table 4) confirmed that pupils find accurate measurement using glass thermometers very difficult, and errors of 1ºC either way are common. Thus, even when free from preconceptions and behaving procedurally correctly, pupils of this age group are unlikely to conclude a scientific answer from this activity based solely on empirical data. Although the activity was used in the current research precisely because of this reason so to enhance error and promote any experimenter expectancy that may have been triggered, this exercise in itself is unsuitable pedagogically because of its inherent inaccuracy. Increasing the resolution of measurements, for example by using a digital thermometer which is easier to read, would be an improvement. Pupils tend not to realise that data collection is inherently random and often does not produce a single definitive 'right answer'; instead a spread of values are collected that approximate to that right answer (Lubben, Campbell, Buffler & Allie, 2001). This is why science curricula promote repeat measurements during experiments. In the current study ideally pupils who had successfully bracketed any preconceptions should have realised this randomness was the reason why they were not measuring exactly the same temperature in each of the three cups.

Overly data-laden experimenting

If a professional scientist blindly accepted unreliable results they would be operating in an overly data-laden manner, applying an inductive scientific method, and would be criticised since the data indicate faulty instrumentation or method. Thus, it could be argued that some pupils in the current study were similarly behaving robotically and did not recognise that their data were indeed inaccurate, so were excessively data-laden. Millar, Lubben, Gott and Duggan (1994) found that 10% of students in their study could not 'see' any patterns in quantitative data, and seemed to take measuring in school science as a kind of mindless ritual, with no clear purpose other than to obtain sets of numbers that actually mean nothing to them. In the current study, since 292 out of 297 reported temperatures were exact to the degree, many pupils were rounding up or down when recording their results, which is scientifically acceptable. But judgments about whether to round up or down were likely to be a source of error that contributed to the wide variety of different temperatures that were recorded. Importantly, because these data were ambiguous in the sense that a choice between two temperatures was possible, it allowed room for bias, either consciously or otherwise.

Accuracy of measurement

The mean temperature range of the random error group is lower than those of the other groups (table 6). The greater spread of temperatures collected by the other groups indicates that factors were at play that made different pupils who underwent the same treatment record values that varied more from the real temperature (although with the whole class phase, for reasons discussed above, it could not assumed that there was a single real temperature). It is proposed that these factors were related to experimenter expectancy. In this sense, the random error group were the best scientists in the sample since they observed with the greatest accuracy. This was due to a complete absence of experimenter expectancy since they had to measure the temperatures of three identical, unlagged cups, and ideally the other pupils should have displayed these same levels of accuracy by bracketing any theory-led expectations and being absolutely objective.

Conditions for the lone experimenters attempted to provide a distraction-free environment for pupils so that accuracy of measurement could be maximised. Since these conditions differed in a number of ways from the whole class, which one or combination of them was specifically causative for the less theory-led behaviour of the lone experimenter pupils? A pupil's temperature range is assumed to be a measure of accuracy, since accurate measurement would involve collecting three equal values (at the 'real' temperature). When the temperature ranges of each phase were compared it was found that pupils in the lone experimenter phase had *not* displayed more accuracy then whole class pupils. Thus, it seems that factors inherent within the lone experimenter conditions that were

designed to improved accuracy of measurement (distraction-free, practice with flashcards, one thermometer per cup) in fact did not impact on accuracy. The random error pupils apart, since all groups had operated at equivalent levels of accuracy this suggests that the absence of peers and/or the close presence of the teacher were governing factors when the lone experimenter T-L pupils successfully differentiated their theory and evidence.

Overly theory-laden mind set

During data collection, 40% of the whole class T-L group not only reproduced the teacher's right answer of fur being the warmest material (as they had also predicted), they also reproduced the rest of their predictions in entirety. The requirement to produce one given result during experimentation appeared to generate a theory-led mindset in these pupils where theory and evidence merged completely. This infers that the problems associated with illustrative practicals may be more extensive than previously demonstrated by other studies, as within a notable proportion of whole class T-L pupils theory-ladeness proliferated to a level beyond that planned by the teacher. Of course, one limitation of the current study is its comparatively small scale, therefore generalisations such as this (based on 8/19 pupils), despite being based on significant statistics, must remain tentative until larger numbers of pupils are surveyed.

Encouraging the acceptance of anomalies

A different problem would be if pupils came to a class with strongly held misconceptions then they would need convincing evidence in order for them to reconstruct an alternative that aligns with accepted science (e.g. author, 2010). Even if they are told the 'right answer', and the experimental phenomenon reliably demonstrates this, any tendency to not differentiate between their own theory (the misconception) and evidence would mean continued existence of that misconception.

A further problem is when the empirical evidence is correct but the textbook theory has been incorrectly applied, e.g. there is some aspect of the system under study that differs from the textbook, as was the case with the T-L treatments in current study. In this case pupils would need to recognise that their results are fine, and then reject that theory, as did the T-L lone experimenters. Illustrative school science sometimes fails to confirm theory because the experiment goes wrong and produces anomalous data. If pupils see anomalies they should be vocal and bring the teacher's attention to them, not blindly discard them and go with the textbook - we do not want them to always fall back on the textbook if there is convincing evidence that says otherwise. If we want pupils to think like scientists, they need to be able to fairly weigh theory and evidence against each other in this way.

The pedagogical usefulness of being overly theory-led

Because the T-L activity in a whole class context generated notable experimenter expectancy, it could be used on its own as a pedagogical tool to teach one aspect of the nature of science. Pupils could carry out the activity and at the end the teacher demonstrates that the temperatures in all three cups are actually equal. This could then be related to the fact that historically, professional scientists have been known to be overly theory-led which has ensued in the collection of biased data that fulfils a self-fulfilling prophesy. This can be at the least, embarrassing for the scientist, and at worst career-ending, so pupils must themselves take note of their empirical data, only rejecting them with good cause and not glibly explaining them away as experimenter error. Alternately, half the class could be told the 'right answer' (T-L) and the other half not told (H-D), then both groups come together at the end for discussion (Millar, 1989).

Implications

Some workers have proposed that school pupils can have a natural tendency to prefer theory over evidence, failing to differentiate between the two entities (Zimmerman, 2007). This tendency could be only exacerbated during illustrative lessons, where theory and evidence *must* correlate. If activities are commonly presented as illustrative exercises then it is possible that procedurally improper behaviours may become part of pupils' repertoires, re-emerging later when they become undergraduates (Birkhead, 2007) or professional scientists (Dunbar, 2000). Although it would be impossible to dispose of the illustrative approach as a way of teaching science concepts (Nott & Smith, 1995), it is recommended that wherever possible, teachers minimise any experimenter effect by presenting practical work as hypothetico-deductive enquiry exercises where the 'right answer' is initially not made obvious to pupils. However, there are well recognised problems with effectively encouraging teachers to deliver enquiry lessons without adequate professional training and support, for instance due to the perceived preparation time required (Blanchard *et al*., 2010).

Data from the current research suggest that an illustrative practical generated less theory-led behaviour when experimenters were allowed to work in private, away from peers, where careful measurement was encouraged in a distraction free environment. From the 1960s-1980s both in the US and the UK, informed by behaviourist theory, modern foreign language teaching in many schools assumed that pupils learn best when carrying out repetitive behaviours in isolated cubicles away from distractions (Roby, 1996). Although language labs have now fallen out of favour, there may be worth in the idea of encouraging science pupils to work more independently, making careful observations with a minimum of disturbance, and then come together with peers at the end in order to compare conclusions. This would be more 'authentic' of professional science, and offer valuable reflective

time. All this said, there was evidence in the current study of some experimenter expectancy when pupils worked in solitary conditions, regardless of the application of the illustrative or the enquiry approach.

Traditionally, pupils work in collaborative groups when they collect data during science lessons (Watson *et al*., 2004). Advantages include peer support for less confident pupils (Gott & Duggan, 1995), opportunities to discuss data interpretation, the fact that children enjoy doing practical work with their friends, and also pragmatic reasons such as a lack of apparatus that exempts solo experimenting. Some writers have highlighted disadvantages of group work. Dominant members take over operations, leaving others redundant (Simon & Jones, 1992). The acquisition of science concepts and process skills are individualistic pursuits, and these are often neglected when working with peers, where the most important thing becomes producing a good joint report (Watson *et al*., 2004).

In Information Technology or Design and Technology lessons, pupils tend to work as individuals on their own projects though still have access to peer and teacher support when needed. Similarly, in high school and undergraduate science, practical work tends to become more individualistic with students having their own sets of apparatus to operate. Future work could investigate the viability of such strategies in the earlier years of school science. Curricula attempt to encourage authentic scientific behaviour in pupils. Perhaps the time has come to look again at how real scientists work, and how best this can be reflected in schools to promote experimentation that is neither overly theory nor data-laden.

References

Alsop, S. (2005). Bridging the Cartesian divide: science education and affect. In S. Alsop (Ed.), Beyond Cartesian Dualism: Encountering Affect in the Teaching and Learning of Science (pp3-16). Dordrecht: Springer.

Asch, S. E. (1951). Effects of group pressure upon the modification distortion of judgments. In H. Guetzkow (Ed.), Groups, leadership and men (pp177–190). Pittsburgh, PA: Carnegie Press.

Atkinson, E. P. (1990). Learning scientific knowledge in the student laboratory. In E. Hegarty-Hazel (Ed.), The student laboratory and the science curriculum (pp119-131). London: Routledge.

Austin, R., Holding, B., Bell, J. & Daniels, S. (1991). Patterns and relationships in school science. Assessment matters: No. 7. London: SEAC/EMU.

Author (2006). Unpublished PhD dissertation.

Author (2007). School Science Review.

Author (2010). Book.

Author (2010). Journal of Research in Science Teaching.

Author (2011). Book chapter.

Author (2011). Research in Science and Technological Education.

Baron, R. S., Vandello, J. A. & Brunsman, B. (1996). The forgotten variable in conformity research: impact of task importance on social influence. Journal of Personality and Social Psychology, 7, 915-927.

Birkhead, T. (2007). 'Let's face it, in terms of real education the school experiment of the past 20 years or so has been a disaster.' The Times Higher Education Supplement, 23rd November.

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A. & Granger, E. M. (2010). Is Inquiry Possible in Light of Accountability?: A Quantitative Comparison of the Relative Effectiveness of Guided Inquiry and Verification Laboratory Instruction. Science Education, 94, 577-616.

Chesterton, G. K. (1933). Saint Thomas Aquinas. Teddington: The Echo Library.

Chinn, C. A. & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. Journal of Research in Science Teaching, 35, 623-654.

Claxton, G. (1986). The alternative conceiver's conceptions. Studies in Science Education, 13, 123-130.

Dean, D. & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. Science Education, 91, 384-397.

Del Carlo, D. I. & Bodner, G. M. (2004). Students' perceptions of academic dishonesty in the chemistry classroom laboratory. Journal of Research in Science Teaching, 41, 47-64.

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. Journal of Developmental Psychology, 21, 49-58.

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

Foulds, K., Gott, R. & Feasey, R. (1992). Investigative work in science. Durham: University of Durham.

Gilovich, T. (1991). How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life. New York: The Free Press.

Gott, R. & Duggan, S. (1995). Investigative Work in the Science Curriculum. Buckingham: The Open University Press.

Greenwald, A., Pratkanis, A., Lieppe, M. & Baumgardner, M. (1986). Under what conditions does theory obstruct research progress? Psychological Review, 93, 216-229.

Gunstone, R. F. & Champagne, A. B. (1990). Promoting conceptual change in the laboratory. In E. Hegarty-Hazel (Ed.), The student laboratory and the science curriculum (pp159-182). London: Routledge.

Gunstone, R. F., Mitchell, I. J. & the Monash Children's Science Group (1988). Two teaching strategies for considering children's science. In The Yearbook of the International Council of Associations of Science Education No. 2, What Research Says to the Teacher (pp1-12).

Hainsworth, M. (1956). The effect of previous knowledge on observation. School Science Review, 37, 234-242.

Hainsworth, M. (1958). An experimental study of observation in school science. School Science Review, 39, 264-276.

Havdala, R. & Ashkenazi, K. (2007). Co-ordination of theory and evidence: effect of epistemological theories on students' laboratory practice. Journal of Research in Science Teaching, 44, 1134-1159.

Hesse, J. J. III. (1987). The costs and benefits of using conceptual change teaching methods: a teacher's perspective. In J. D. Novak (Ed.), Proceedings of the Second International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Vol. 2 (pp194-209). Ithaca, NY: Cornell University.

Keiler, L. S. & Woolnough, B. E. (2002). Practical work in school science: the dominance of assessment. School Science Review, 83, 83-88.

Kuhn, D., Amsel, E. & O'Loughlan, M. (1988). The development of scientific thinking skills. San Diego: Academic Press.

Kuhn, D. & Pearsall, S. (2000). Developmental origins of scientific thinking. Journal of Cognition and Development, 1, 113–129.

Lawson, A. E. (2010). Basic inferences of scientific reasoning, argumentation, and discovery. Science Education, 94, 336-364.

Lubben, F., Campbell, B., Buffler, A. & Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshmen. Science Education 85, 311-327.

Lubben, F. & Millar, R. (1996). Children's ideas about the reliability of experimental data. International Journal of Science Education, 18, 955-968.

Martindale, D. A. (2005). Confirmatory bias and confirmatory distortion. Journal of Child Custody, 2, 31-48.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R. *et al*. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. Journal of Research in Science Teaching, 41, 1063 – 1080.

McCormas, W. F. M. & Moore, L. S. (2001). The expectancy effect in secondary school biology instruction: issues and opportunities. The American Biology Teacher, 63, 246-252.

Mercer, N., Dawes, L. & Staarman, J. K. (2009). Dialogic teaching in the primary science classroom. Language and Education, 23, 353-369.

Millar, R. (1989). Bending the evidence: the relationship between theory and experiment in science education, In R. Millar (Ed.), Doing Science: Images of Science in Science Education (pp38-61). London: Falmer Press.

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, 207-248.

Nott, M. & Smith, R. (1995). Talking your way out of it, "rigging" and "conjuring": what science teachers do when practicals go wrong. International Journal of Science Education, 17, 399-410.

Poletiek, F. (2001). Hypothesis-Testing Behaviour. (Essays in Cognitive Psychology). Hove: Psychology Press.

Preist, S. (2007). The British Empiricists (2nd Edition). Abingdon: Routledge.

Rigano, D. L. & Richie, S. M. (1995). Student disclosures of fraudulent practice in school laboratories. Research in Science Education, 25, 353-363.

Roby, W. B. (1996). Technology in the service of foreign language teaching: The case of the language laboratory. In D. Jonassen (Ed.), Handbook of Research on Educational Communications and Technology, 2nd ed (pp523-541).

Rogan, J. & Aldous, C. (2005). Relationships between the constructs of a theory of curriculum implementation. Journal of Research in Science Teaching, 42, 313-336.

Rosenthal, R. (1966). Experimenter Effects in Behavioral Research. New York: Appleton-Century-Crofts.

Simon, S. A. & Jones, A. T. (1992). Open Work in Science: A Review of Existing Practice. London: Kings College.

Stangor, C. (2004). Social groups in action and interaction. New York, NY: Psychology Press.

Tiberghien, A., Sere, M. G., Barboux, M. & Chomat, A. (1983). Etude des representations prealables de quelques notions de physique et leur evolution. Rapport de recherche, LIRESPT, University of Paris VII, Paris.

Watson, J. R., Swain, J. R. L. & McRobbie, C. (2004). Students' discussions in practical scientific inquiries. International Journal of Science Education, 26, 25-45.

Wellington, J. (1981). 'What's supposed to happen, sir?': some problems with discovery learning. School Science Review, 63, 167-173.

Wigal, J. K., Stout, C., Kotses, H., Creer, T. L., Fogle, K., Gayhart, L. & Hatala, J. (1997). Experimenter Expectancy in Resistance to Respiratory Air Flow. Psychosomatic Medicine, 59, 318-322.

Wood, W. (2000). Attitude change: Persuasion and social influence. Annual Review of Psychology, 51, 539-570.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. Developmental Review, 27, 172-223.

Zimmerman, C., Raghavan, K. & Sartoris, M. L. (2003). The impact of the MARS curriculum on students' ability to coordinate theory and evidence. International Journal of Science Education, 25, 1247-1271.

[i] This is sometimes referred to in the literature as the *observer expectancy effect*.

[ii] Trialling using both glass thermometers and datalogger temperature probes had shown that water in the three cups remained at the equivalent temperatures over long periods.

[iii] The water used in the experiment was at ambient temperature. However, if one were to dip one's fingers into this water it would feel colder than the surrounding air because heat is transferred away from the fingers more quickly. In wintertime, the sea may be at a higher temperature than the surrounding air but a person would lose body heat far more rapidly if they fell into the sea, compared to being on dry land, due to this enhanced heat transfer effect.

[iv] When numerical counts are cited they are exclusive of uncodeable responses.

[v] Note that all claims in the subsequent sections of this article are based on statistically significant associations, unless otherwise stated.