Improving the Validity of Contextualised Questions

Ayesha Ahmed and Alastair Pollitt
University of Cambridge Local Examinations Syndicate

Paper to be presented at the BERA Conference, Leeds, September 2001.

The opinions expressed in this paper are those of the authors and should not be taken as official policy of the University of Cambridge Local Examinations Syndicate or any of its subsidiaries.

Contact details

Ayesha Ahmed, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU
ahmed.a@ucles.org.uk

Improving the Validity of Contextualised Questions

Ayesha Ahmed and Alastair Pollitt

Abstract

Setting science exam questions in real-world contexts is widespread. However, when students are reading contextualised questions the cognitive processes provoked by the context can interfere with their understanding of the science in the question. Validity is then compromised in the sense that a question is only valid if the students' minds are doing 'what we want them to show us they can do'. The *focus of a context* for a question is the extent to which the most salient aspects of the context correspond to the main issues addressed in the question. Can we improve the validity of contextualised questions by setting them in more focused contexts? A focused context will help activate relevant concepts, rather than interfering with comprehension and scientific thinking. In this study, questions in a Key Stage 3 science test were manipulated to alter the focus of their context. We report the effects of these changes on the difficulty and validity of the questions and, with additional evidence from interview protocols, we consider the effects of context *focus* on the question answering process.

Context and Validity

We define the validity of exam questions in terms of cognitive psychology, that is in terms of the processes occurring in students' minds when they are answering the questions. We are trying to measure students' understanding or skills using a particular question, so we want this question to cause certain processes to occur in their minds. These processes in turn cause the 'things we want the students to show us they can do'. We therefore say that a question can only be valid if *'the students' minds are doing the things we want them to show us they can do'*.

This clearly requires the question writer to exert a level of control over the students' minds, and their instrument for doing this is the question. The question should cause the right processes to occur in the minds of all the students who answer it although some will carry out these processes better than others. This enables examiners to use students' answers to a question as evidence for how well they can do 'the things we want them to show us they can do'; that is it enables examiners to measure the students validly.

Any aspects of a question that get in the way of the students 'doing the things we want them to show us they can do' are a threat to validity. Anything that reduces the examiners' level of control over the processes occurring in students' minds when they answer the questions will get in the way of measuring what we want to measure. If we don't know what processes a question is going to provoke in the students' minds then we cannot say that we are measuring their understanding of that topic or their knowledge about it or their skills.

Setting questions in real world contexts is one potential threat to validity. The effects of a real world context on the processes that occur in students' minds when they are answering a question are in some ways unpredictable: a context will have different effects on different students since it will differ in familiarity to them. It is therefore much more difficult for examiners to be in control with a contextualised question, and much harder to say we are measuring understanding of a particular topic. In spite of these worries, questions are often set in real world contexts, and there are good reasons for doing this.

Why use Context?

Those involved in the assessment of mathematics and science often wish to assess whether or not students can apply their mathematics or their science to new situations, particularly real-life ones. They are interested in whether students can *do* the science or the maths rather than just *understand* it, with the assumption that if students can apply the concepts in this way then they must have understood them. For this reason, mathematics and science exam questions are often set in real world contexts. This is very popular at GCSE and at Key Stages 2 and 3.

Setting a question in context can also make it more concrete, so that it is about a specific example rather than an abstract concept. The assumption is that this makes a problem less demanding. As well as aiming to assess students' ability to apply knowledge, question writers use context to try to make the science or the maths seem relevant to everyday life in order to motivate students and keep their interest. These are important considerations which can be addressed during teaching. It may not be necessary or even advisable to try to address these aims during written paper assessment. The assessment should be relevant to the teaching but it is more important that the assessment is valid in the sense that the students' minds are doing 'the things we want them to show us they can do'. A good context might allow us to measure students' ability to apply knowledge, but a bad context can prevent us from measuring anything at all.

The aim of this paper is to show how the degree of focus of a question's context can affect students' performance and to discuss the possibility that some contexts are better than others for ensuring validity. The idea of focus is discussed below.

Why Context is Dangerous

The assumption that context makes a question less demanding cannot be justified as the context itself adds extra demands. Three sorts of demands may be considered, as possible reasons why context can threaten validity.

Language
Putting questions into context inevitably involves using extra words to ask the question. If students have to read more text in order to answer a question then their reading ability is being tested as well as their understanding of science. At Key Stage 3 there are students with poor reading skills and we would not wish to penalise them for this in a science exam. The sort of language used in a real world context is often more complicated than the language needed to ask about the science, involving metaphors or even culturally specific terms. The terminology of the context may involve vocabulary relating to adult concepts; Boaler (1993) pointed out the importance of avoiding 'adult metaphors' such as wage slips.

Familiarity
Inevitably any real world context will be more familiar to some students than others. Those who know a lot about the context must select aspects that are relevant and ignore everything else they know about it. They might have difficulty deciding whether to use their everyday knowledge of the context to answer the question, or just to stick to the science. Those who are unfamiliar with the context might be put off attempting the question, and knowing certain aspects of the context might even be necessary for answering the question. Evidence from a previous study (Ahmed & Pollitt, 2000) shows that students sometimes confuse context with content. They think that the question is asking about a part of science that they have not learned, whereas in fact it is familiar science in a new and unfamiliar context.

Attention
Contextualised questions necessarily contain a great deal of irrelevant information, requiring students to select what is relevant to answer the question. If we do not want to assess this skill

as part of science then the amount of irrelevant information should be kept to a minimum. Even pictures can cause difficulty as they are often full of irrelevant information which, because of their high salience, is particularly distracting. Attentional resources are limited, and can easily be distracted and divided if students have to deal with aspects other than the science in the question. Students are under stress during an exam so their attentional capacities are already reduced.

There has been some research on the effects of context in maths questions but very little in science. It has been demonstrated that problems in context can be more difficult to solve than solving context-free problems. Kotovsky, Hayes & Simon (1985) investigated performance on the Tower of Hanoi problem-solving task in a *sparse* versus an *enriched* context, and found that the latter was more difficult. Mevarech and Stern (1997) reported that both 11 year-olds and adults found it easier to solve tasks embedded in a sparse context than those embedded in a real world context. They looked at the interpretation of linear graphs and found that those participants who were given the graph within a sparse context used abstract logic-mathematics to solve the task. Those who were presented the graph within a real context used different non-mathematical knowledge structures relating to practical reasoning and everyday knowledge to interpret the graph, and did less well on the task. They suggest that real contexts not only divert students' attention from the mathematical task but can activate simplistic models rather than abstract thinking.

This distinction is supported by considerable evidence for two forms of rationality in human reasoning. Sloman (1996) described an *associative system* which is typically employed in everyday life, operating on the basis of regularities observed in everyday life, and a *rule-based system* which is used to reason logically and scientifically. Evans & Over (1996) came to similar conclusions, and emphasised how these two systems interact and interfere. If the effect of contextualisation is that students' minds are not doing scientific thinking this would seem, by our definition of validity, to invalidate the test.

Boaler (1993) found that students tend to choose the procedure with which to do a mathematics problem according to the context of the task rather than the task itself, and Nickson (1998) found that students tend to focus on a single element of a mathematics question, for example the price of an article that has been illustrated, and carry this through to the end of the task. When questions were put into context, the aspect of the question on which students chose to focus was more likely to be a contextual one, often resulting in wrong answers.

Mevarech and Stern (1997) also found that experience gained in the sparse context task could be transferred to the real context task, but not vice versa. Anderson, Reder & Simon (unpublished) argue that if mathematics is learnt in context it is more difficult for students to transfer their knowledge, whereas mathematics that is taught in the abstract is more successfully transferred. Authentic problems in which mathematics applies to everyday life can be encountered after students have learnt the cognitive skills necessary to solve those kinds of problems. If they have learnt the underlying cognitive skills in an abstract form, they will be able to transfer them to any real world task.

We can expect that these issues will apply to contextualised science questions in a similar way.

The Question Answering Process

In order to discuss what we mean by a focused context it is important to consider the process by which students form an understanding of a question or task. We have developed a six-phase model of the question answering process (Pollitt & Ahmed, 1999) that is an attempt to describe the psychological processes that occur when students answer exam questions.

The model starts with Learning before the exam, followed by the first phase during the exam – Reading the question. When reading a question students form a mental representation of what the question is asking them to do. Each student forms their own personal idiosyncratic representation provoked by the text and mostly built from pre-existing knowledge. The words in the question are "cues to build a familiar mental model" (Johnson-Laird, 1981). As students are forming this representation, certain concepts are activated in their minds. Many irrelevant as well as relevant concepts will be activated in the students' minds, and they have to select which to use to answer the question. This automatic activation process happens before the reading of the text reaches consciousness so the student will not be aware of all the concepts that have been activated. However all of these concepts can affect the subsequent interpretation of the task. Evans (1989) calls this 'pre-attentive bias' – the students' interpretation of the question is biased by linguistic processing that occurs before conscious attention to the task.

The next three phases of the model are Searching – the student searches the network of activated concepts, Matching – a relevant concept is identified and matched with their understanding of the question, and Generating – an idea of an answer is generated. The final phase is Writing – the student turns their mental representation of the answer into - usually - a string of words.

Context will have its main effect during the Reading phase, causing potential misunderstandings while students build mental representations of the question that are not what the examiner intended. A contextualised question will activate many more irrelevant concepts than a pure science question and this will also affect the Searching, Matching and Generating phases of the model. This is best thought of in terms of schemas (Bartlett, 1932), pre-existing structures within which concepts are activated. A real world context will activate familiar schemas which create powerful expectations of what the question is about. These may well be inconsistent with or interfere with the schemas provoked by the science in the question, that is the schemas the examiners intended to provoke in the students' minds.

A Focused Context

A context that is *focused* will be less likely to cause students to form a different understanding of the question from the one the examiner intended. We define *focus* as the extent to which the most salient aspects of the context correspond to the main issues addressed in the question. A focused context helps to activate relevant concepts rather than interfering with comprehension and scientific thinking. Any context will bias students towards certain expectations of what the question is asking; it will bias them to think about the question in a particular way. A focused context will bias students to think about the question in the right way, that is it will get them thinking about the right bit of science in the right way. It does this by facilitating the activation of the same schemas that are provoked by the science in the question. If this is the case then the concepts activated are those that are needed to answer the question.

This does not necessarily make the question easier for students who don't know the science. It may well make it easier to get the mark, not because the science itself is any easier, but because misunderstandings of the question are less likely. The scientific content of the question is intended to allow examiners to measure how well students understand or can do the science, and a context should not hinder this process. We want to cause all of the students to be trying to answer the same question, the intended question. Only then can we assess how well they are doing it. If some of them are answering a different question - the one they think they are being asked because the context suggests it - then we can't measure these students validly. We have lost control over what we are measuring.

We have found many examples of questions which caused this kind of misunderstanding (Ahmed & Pollitt, 2000; Pollitt & Ahmed, 2000). One of these was a question about a chemical reaction set in the context of a golf ball dissolving in the sea. The student we observed and interviewed thought about the question for a long time before saying '*I haven't really done it*.' He was referring to not having studied what happens to golf balls made of this substance when they are in sea water. He had learned the necessary chemistry to answer the question but did not realise this as the context was unfamiliar. In this case he mistook context for scientific content. Another example was a question about the disadvantages of using oil to remove old paint. One student said: *'It sticks and it's messy – I didn't know if that was scientific or not – whether to use common sense or just scientific principles.'* The real world nature of the context was causing her to hesitate over how to answer.

If the context is focused on the science at issue then students are less likely to form a very different idea from the examiner of what the question is asking. The natural focus of the context should correspond to the main point of the question. For example a question set in a shopping context with a number of goods at different prices should involve students calculating the total bill, as this is the natural focus of a shopping context, central to our 'shopping' schema.

All parts of a question should lead towards the focus and all should be equally related to the context. Sometimes questions are written in which some parts relate more strongly to the context than others. This should be avoided as students will expect the context to be relevant to every part of the question and may well answer in terms of the context even if that is not what the examiner intended. If a context is well focused and has a purpose – to assess the application of the science - rather than just being a setting to ask about some science, then it will be much easier to write question parts that relate to it. Wiliam (1997) points out that mathematics is often taught in contexts which are not related to the subject matter being taught. These contexts are designed to motivate learners by making the task seem relevant. He calls this 'maths looking for somewhere to happen'. This applies equally to science, and when translated to assessment, it results in students having to decipher what the examiner is intending them to do by decontextualising the task.

For this study questions were chosen that offered opportunities to write different versions in contexts that were different in focus. In each case one version is considered to be *unfocused*, in that the point of the question does not correspond to the most salient features of the context. Other versions were written to be more focused; each example is described separately below. We looked for changes in the difficulty of the question to show that context does change students' question answering processes.

We also examined how questions with more or less focused contexts correlated with the rest of the test. This addresses the issue of whether the changes in context make the questions better at testing the understanding and skills that we want to measure, rather than just easier or harder. We are interested in whether a context that seems to be natural and focused for the question we want to ask is going to cause the right processes to occur in the students' minds and therefore result in a valid question.

The purpose of the research is to understand how contexts and questions interact, and how focus works. If we succeed we will be able to assess the application of scientific understanding, using contextualised written examination questions, without threatening the validity of the test.

Method

Questions
Three questions that were originally developed for Key Stage 3 Science but never used in a live test were adapted to produce either three or four versions so that the focus of the context was varied. A single maths question was also adapted to produce two different versions. Three further questions were chosen from past Key Stage 3 Science papers to act as control questions. These questions were not adapted and were common for all students, to allow equating of the papers as described below. The final question papers consisted of seven questions: six Science and one Maths. Given the substantial amount of calculation in Science tests at this level it seemed worth including a pure arithmetic question in what is otherwise a science test. The questions can be found in Appendix 1.

Participants
The papers were taken by 405 Year 9 students taken from the whole ability range in two local comprehensive schools. The test was carried out during February and March, that is two or three months before they took their Key Stage 3 tests. By this time they had covered all the National Curriculum content necessary to answer our questions.

Design
Twelve different forms of the test were made up so that each version of a manipulated question occurred in a paper with every version of the other questions. This can be seen in the table in Appendix 2. Students were given the tests in groups of about 30 during their Science lessons. The twelve forms were distributed systematically amongst students in each class so that approximately the same number of students did each one. This also ensured that the forms were equally distributed to the more able and less able students.

Procedure
Students were given 30 minutes to attempt all of the questions on the paper. After this the question papers were collected and marked. A total of fourteen students were interviewed immediately after completing the test. They were interviewed in pairs and asked a series of questions which can be found in Appendix 4.

Results

Analysis
Data from the twelve separate test forms were combined into a single data set and Rasch analysed. In each form there were 13 common marks, and 12, 13 or 14 marks in the experimental question versions. In the analysis the common questions serve to fix a single scale on which difficulty parameter estimates for all the questions, common and specific, can be expressed. In practical item banking it is normal to equate tests using an overlap of less than 20% of the total available marks; the 50% overlap we used here should ensure that the equating was sufficiently dependable to support the research analysis.

Rasch difficulty parameters are conventionally fixed in such a way that the mean of all the parameters is 0, and this convention is followed below. Our interest here is in the comparison between the difficulty parameters of different versions of the same question. If one version is easier than another its difficulty parameter will be lower than the other's; often this will mean that it is more negative rather than less positive.

In a Rasch context a misfit statistic is normally used to diagnose poor item quality. For our purposes - again because we are interested only in comparing different versions of the same question - it was judged simpler to use a more traditional index, the point biserial correlation between the item and the rest of the test taken by that student. That the maximum score varied from 25 to 27 on different test forms will perturb this correlation, but only a little, and the way

that the question versions were combined in the twelve forms ensures that there is no bias in the correlations. We can then assume that higher correlation means a more valid question; experience suggests that, for a test of this length, a correlation below 0.2 would indicate that the question is insufficiently valid.

The student samples

The table below shows the result of an analysis of variance comparing the mean ability estimates of the groups who took each test form. There was no significant variation between the twelve groups.

One Factor ANOVA X : Ability by :Test form

Analysis of Variance Table

Source:	DF:	Sum Squares:	Mean Square:	F-test:
Between groups	11	3.282	.298	.262
Within groups	393	447.248	1.138	p = .9919
Total	404	450.53		

Question 1 – Tea, Cooling and Warming

The original question was Version 1 – Tea. The idea of an experiment about making tea in which the darkness is measured is not a very focused context as it is not a common experiment in school science. It is also difficult to imagine how Becky would *measure* how dark the tea was. The correct curve in the graph is A and this is the sort of curve that would result from an experiment where water is cooling, so it should activate a 'cooling curve' schema in students' minds. The most focused context for this graph would therefore be an experiment relating to a cooling curve. Version 2 was therefore written as an experiment about cooling liquids, which is a familiar experiment in school science. This was reflected in the interviews, and one student said '*Question 1 (Version 2, Cooling) was quite simple cos we'd done a lot of work on cooling curves.*' One concern about this version was that the position of the labels on the graph would be counter-intuitive with 'Cold' at the top and 'Warm' at the bottom. In Version 3 the context was switched to warming liquids so that the labels on the graph could be reversed.

Question 1a	Version 1 Tea	Version 2 Cooling	Version 3 Warming
Focus	Not focused	Very Focused	Focused
Difficulty	0.777	1.146	1.000

Correlation	0.128	0.054	0.158

The point biserials for all three versions were very low, leading us to conclude that there was something fundamentally wrong with the question. The differences in the difficulty estimates also are not statistically significant. Since the question seems essentially invalid it is not appropriate to try to interpret such small changes in a validity study. Question 1 was dropped from consideration.

The difficulty values for this and all other questions can be seen in graphical form in Appendix 3.

Questions 2 and 3 were common questions and were not manipulated.

Question 4 – Coca-cola and Ski Pass

Version 1 was about Coca-cola and told students how much a crate of 12 cans cost. Then instead of the natural focus of the context which would be a division to calculate how much one can costs, the students were required to do a multiplication to find out how much 7 crates cost. The fact that there were 12 cans in a crate was irrelevant information, and the natural focus of the context did not match with the natural focus of the maths. One student said she had started the Coca-cola question by working out the cost of 7 cans but had then realised this was not the question and crossed it out: '*I just got really muddled u*p'. The real world schema that is activated for this question involves working out how much you would have to pay to buy a certain number of cans from the crate.

Version 2 was the more focused version. In this case the same multiplication had to be carried out but this time the focus of the context matched the natural focus of the maths. There is a clear reason why you would want to do the multiplication to work out how much a ski pass costs for 7 days. The schema for a ski pass includes the idea of paying for a week's skiing.

	Version 1 Coca-Cola	**Version 2 Ski Pass**
Focus	Not focused	Focused
Difficulty	0.344	-1.326
Correlation	0.036	0.292

Version 2, the ski pass, was considerably easier than Version 1. The more focused context of the ski pass caused the same calculation to be easier than in the Coca-cola context in which the focus was division. As well as being easier, the ski pass question also correlated much better with the rest of the test compared to the coca-cola question. This high correlation justifies the inclusion of such a question in the test. It also indicates that the ski pass version was a more valid question, measuring what we were trying to measure. The better able students were more likely to get it right and the least able students were more likely to get it wrong.

Question 5 was a common question and was not manipulated.

Question 6 – Lolly on balance, Lolly on electric, Ice on electric and Context-free ice

The unfocused version of this question was set in the context of an ice lolly on one pan of a balance. This is not a very natural context as it is unlikely that someone would use an ice lolly in this way, just letting it melt in an experiment instead of eating it, and the idea of doing this will cause dissonance for some students. It is more likely that a block of ice would be used for this kind of experiment.

The scientific idea behind the question is for students to decide what happens to the mass of an object when it changes state. This is the focus of the science. What the students actually have to do is look at the picture of the ice lolly on a balanced scale pan, decide what happens to the mass of the ice lolly when it melts, and then decide how this affects the scale pan. They are given possibilities of 'up', 'down' and 'the same level'. The focus of the context is the ice lolly on the scales, and the dominant schema that is activated will be about melting ice lollies, with associated focus on waste and loss, and not the scientific schema of changes of state. The picture of the balance is also a potential cause of confusion to students. It looks as though the liquid from the melted ice lolly would overflow the scale pan, and this of course would change the correct answer from 'the same level' to 'up'. Of course 'up' corresponds to the mass going 'down' which is another potential cause of confusion.

Version 2 was written so that the ice lolly was on electric scales in a pan which would not overflow. It is also more likely that electric scales would be used in this sort of experiment. The digital reading on the electric scales avoided the extra inference that students had to make in Version 1 about whether the pans would move and which way. Version 3 used a block of ice instead of an ice lolly, again a more likely scenario in a science experiment, and also avoiding the interference of the lolly stick. The context of doing an experiment with ice in a lab and leaving it over the weekend is much more natural than melting an ice lolly and then leaving it for two days. The focus now is also on state change of water and evaporation, rather than on ice lollies as in Versions 1 and 2.

Version 4 was written with no context – the science underlying the other versions was asked directly in this version. The no-context version also avoided another problem: when students are asked a question in which there is a starting state, an event and then an end state they are tempted to say that the end state is different from the start state whatever the intervening event might be (Donaldson, 1978). The first three versions work like this with a picture of the starting situation, a description of an event, and then a question about the scales afterwards. The focus is on some kind of change in these versions. In Version 4 the question was phrased differently because it was out of context, and there was no picture, so students were less likely to be tempted to say there had been a change. The students were not asked to choose whether or not the reading had changed but instead were asked simply to decide on the mass of the water.

Question 6a	**Version 1 Lolly balance**	**Version 2 Lolly electric**	**Version 3 Ice electric**	**Version 4 Ice**
Focus	Not focused	Focused	Very Focused	Context-free
Difficulty	1.053	0.934	0.918	0.450
Correlation	0.207	0.296	0.358	0.285

Question 6(a), deciding on the scale reading, was easiest in the context-free version and most difficult in the unfocused 'balance' version. The correlation with the rest of the test was lowest in the unfocused version and highest in the very focused one.

Question 6b(i)	**Version 1 Lolly balance**	**Version 2 Lolly electric**	**Version 3 Ice electric**
Focus	Not focused	Focused	Very Focused
Difficulty	-0.376	-0.929	-1.404
Correlation	0.251	0.438	0.434

Question 6(b)(i), how the scale reading changes after 2 days, appeared only in Versions 1, 2 and 3. It was omitted from Version 4 because it could not easily be asked out of context. Version 3, the very focused version, was the easiest and Version 1 the most difficult. The

correlations for Versions 2 and 3 were substantially higher than that for Version 1, so again the more focused versions were easier and better correlated with the rest of the test.

Question 6b(ii)	Version 1 Lolly balance	Version 2 Lolly electric	Version 3 Ice electric	Version 4 Ice
Focus	Not focused	Focused	Very Focused	Context-free
Difficulty	-1.812	-1.994	-2.309	-1.798
Correlation	0.307	0.323	0.385	0.387

Question 6(b)(ii) proved to be an easy question in all four versions, with Version 3 being the easiest. In Version 4 students are not led to the idea of evaporation as they are in the other three versions via part (b)(i). The focused context in Version 3 helps the students to think of evaporation. The correlations again showed the highest correlation for Versions 3 and 4 and the lowest for Version 1.

Question 7 – Alligators, Seeds and Threads

The unfocused version of this question was about alligator eggs, and how temperature influences the sex of hatchlings. There were a number of aspects of the context of this question that we thought could cause confusion. One issue was that the graph has two vertical scales, one of which has zero at the top, and students may not have seen this kind of graph before. The idea of the science in the question was to test students' ability to read and interpret graphs; in particular the question was about a graph that showed two mutually exclusive outcomes, in this case males or females.

We wrote two more versions that tested the same graph skills using more focused contexts. Version 2 was about seeds that either germinated or did not germinate depending on temperature. This is likely to be a more familiar context than alligator eggs hatching as male or female. The idea of seeds germinating is also more relevant to the students and they should be able to relate it to their science lessons. They are likely to have a schema for seeds germinating that includes temperature affecting germination.

The context used in Version 3 was again a focused context that students could relate to their physics lessons. Unlike Version 2, it also allowed for two vertical scales on the graph to match the original version.

Question 7a	Version 1 Alligators	Version 2 Seeds	Version 3 Threads
Focus	Not focused	Focused	Focused
Difficulty	0.819	-0.574	0.446
Correlation	0.427	0.505	0.556

In question 7(a), describing the graph, the seeds version was the easiest and the unfocused one the most difficult. Both of the focused versions proved more valid than the unfocused one.

Question 7b	Version 1 Alligators	Version 2 Seeds	Version 3 Threads
Focus	Not focused	Focused	Focused
Difficulty	-0.404	-0.997	-0.624
Correlation	0.278	0.354	0.475

Question 7(b), which involved reading the graph, was easiest in the seeds version where the graph only had one vertical scale. We could not use two vertical scales for this version as it would not have been appropriate within the context. One of the students who was interviewed commented '*The graph on her one is easier*.'(Seeds). Another commented about the reverse scale on the graph in the original version: '*I had to cross it out because I looked at the graph*

and realised that they were the opposite way round….the two 100s are the wrong way up.'
Again the focused versions proved more valid than the unfocused one.

Question 7c	Version 1 Alligators	Version 2 Seeds	Version 3 Threads
Focus	Not focused	Focused	Focused
Difficulty	-0.887	1.425	-0.673
Correlation	0.400	0.410	0.623

Question 7(c) was also a graph reading question but in this case the seeds version was much more difficult than the other two versions. This question contained a negative, '**FAIL** to germinate' and although it was typed in bold capitals some students answered the question 'At what temperature did 20% of the seeds germinate?'. This may have been because they misread the question by either missing the word 'fail' or just seeing the 20% and looking for 20 on the graph instead of 80. Some students may not have understood how a graph should be read and were not able to work out that they needed to subtract 20 from 100 and then read across from 80 on the graph. As this is the skill that is being tested this would be a valid way of getting no marks, unlike the other invalid ways of going wrong.

Question 7d	Version 1 Alligators	Version 2 Seeds	Version 3 Threads
Focus	Not focused	Focused	Focused
Difficulty	0.386	-0.175	-0.055
Correlation	0.411	0.479	0.439

Question 7(d) asked how the number of seeds germinated, or the number of females, or the strength of the thread, could be maximised. Again the unfocused version was the most difficult and the seeds the easiest, and the focused versions were more valid.

Question 7e	Version 1 Alligators	Version 2 Seeds	Version 3 Threads
Focus	Not focused	Focused	Focused
Difficulty	1.072	0.103	1.871
Correlation	0.140	0.249	0.331

Lastly, question 7(e) asked about what would happen at 50^{o}C or 50N. Here the most difficult was the threads version and the easiest was the seeds – both focused. There is a problem, though, with the threads question in that it was not possible to ask it in a form parallel to the other versions, and the comparison may not be fair. In this case the unfocused question correlates so poorly with the rest of the test that it might be considered invalid; the focused ones were much better.

Conclusions

In every case the questions used in this study were improved when they were put into more natural and focused contexts. That is they were better questions when contexts were created that were designed to provoke the same schemas in the students' minds as the science or maths in the question. These were 'better questions' in two ways. First, focused contexts made it easier for students to get the mark. The underlying maths or science was not changed, so what was making it easier was that the context caused fewer misunderstandings of the question. Secondly these questions were better because performance on these correlated well with the rest of the test. This means that the questions with more focused contexts were better at measuring what we were trying to measure with the overall test, so they were more valid questions. Students who were more able and did well overall on the test were more likely to get these questions right, and students who did badly overall on the test were more likely to get them wrong.

Examiners should aim to create a question which naturally fits a real world context so that the issues central to the science in the question are also central to the context. If this is achieved then similar schemas will be activated in the students' minds by the science and the context so that there is no conflict or interference from the context when the student is forming a mental representation of the task. The context should not just be a real world setting used to ask about the science; in particular context should not be used to provide an 'interesting' setting for the task as this often means a very unfocused and sometimes bizarre context. Instead the context should be an integral part of the question.

If this cannot be achieved for a particular aspect of science or a particular topic then it may be best to address the issues of application of this topic to the real world in the teaching of the subject but not in the formal written assessment. Otherwise we run the risk of causing such misunderstandings of the task so that we cannot measure students' achievement on this topic at all. Another possibility in this situation is that suggested by Brown (1999). She argues that mathematics exam questions should concentrate on the mathematics that is the essence of the school subject, not on the real world, and that the place for contextualised tasks is in project work. When students carry out projects they have much more of an opportunity to explore the complexities of applying their maths to the real world and can do so in a relaxed environment with guidance from the teacher.

In order to ensure validity we need to know that 'the students' minds are doing the things we want them to show us they can do'. The best way to ensure this is to make questions as free as possible from anything that could cause comprehension difficulties that prevent students from carrying out the task. Context, unless it is focused, can cause such comprehension problems and is therefore a threat to validity. As one student said of the non-focused contextualised questions, '*…the way the questions were asked made you think you didn't know it'*.

References

Ahmed, A. & Pollitt, A. (2000) Observing Context in Action. *Paper presented at IAEA conference, Jerusalem, May 2000.*

Anderson, J.R., Reder, L.M. & Simon, H.A. (unpublished paper) Applications and Misapplications of Cognitive Psychology to Mathematics Education. *Unpublished paper at http://act.psy.cmu.edu.*

Bartlett, F.C. (1932) *Remembering*. Cambridge: Cambridge University Press.

Boaler, J. (1993) When Do Girls Prefer Football to Fashion? An analysis of female underachievement in relation to 'realistic' mathematics contexts. *British Educational Research Journal, 20 (5) 551 – 564.*

Brown, M. (1999) One Mathematics for All? In Hoyles. C., Morgan, M. & Woodhouse, G. (Eds.) *Rethinking the Mathematics Curriculum.* Studies in Mathematics Education Series: 10. Falmer Press, London, pp 78-89.

Donaldson, M. (1978) Children's Minds. Fontana.

Evans, J. St. B.T. (1989) *Bias in Human Reasoning: Causes and Consequences*. Hove: Lawrence Erlbaum.

Evans, J. St. B.T. & Over, D.E. (1996) *Rationality and Reasoning*. Hove: Lawrence Erlbaum.

Johnson-Laird, P.N. (1981) Mental Models of Meaning. In Joshi, A.K., Webber, B.L. & Sag, I.A. (Eds.) *Elements of Discourse Understanding*. Cambridge: Cambridge University Press.

Kotovsky, K., Hayes, J.R., & Simon, H.A. (1985) Why are Some Problems Hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17, 248-294.*

Mevarech, Z.R. & Stern, A. (1997) Interaction Between Knowledge and Contexts on Understanding Abstract Mathematical Concepts. *Journal of Experimental Child Psychology, 65, 68-95.*

Nickson, M. (1998) What is the difference between a pizza and a relay race? The role of context in assessing KS2 Mathematics. *British Journal of Curriculum & Assessment, 7 (3), 19-23.*

Pollitt, A. & Ahmed, A. (2000) Comprehension Failures in Educational Assessment. *Paper presented at ECER conference, Edinburgh, September 2000.*

Pollitt, A. & Ahmed, A. (1999) A New Model of the Question Answering Process. *Paper presented at IAEA conference, Slovenia, May 1999.*

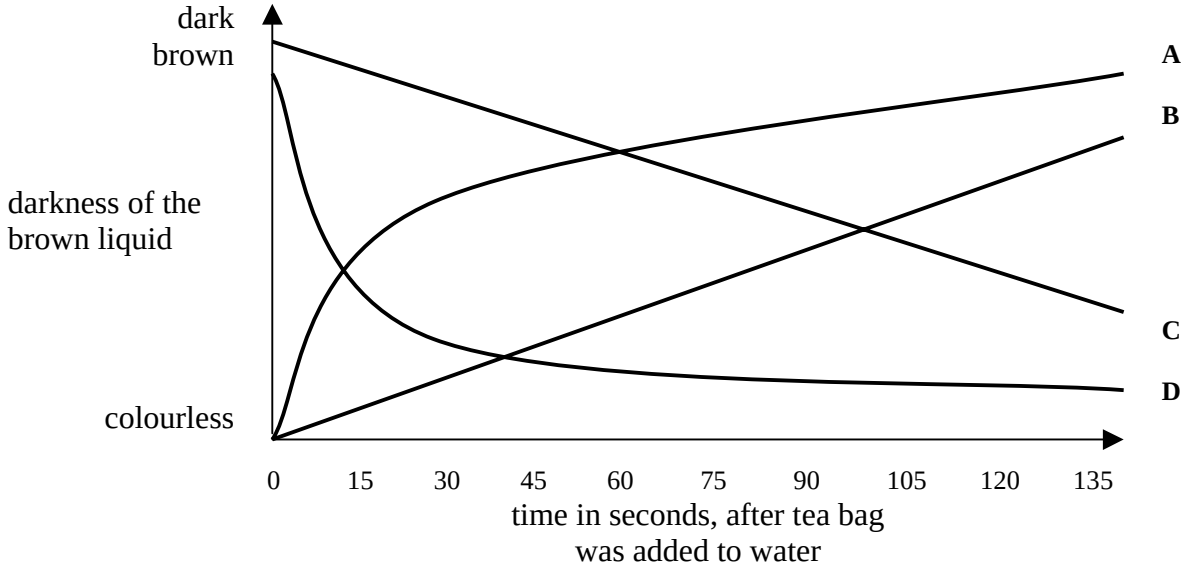
Sloman, S.A. (1996) The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin,* 119, 3-22.

Wiliam, D. (1997) Relevance as a MacGuffin in Mathematics Education. *Paper presented at British Educational Research Association Conference, York, September 1997.*

Appendix 1 – Questions

(Version 1)

1. In an experiment about making tea Becky put one tea bag in a beaker and added 50 cm³ of warm water. She stirred the liquid slowly. Every 15 seconds she took out 2 cm³ of the liquid and measured how dark it was.



(a) Which graph, A, B, C or D shows how the colour of the liquid changed?

_____ *1 mark*

(b) (i) Becky took out 2 cm³ samples of the liquid each time. Why must she always put the sample back after she has tested it?

_____

_____

1 mark

(ii) What piece of apparatus could she use to measure the volume of the 2 cm³ samples of liquid?

_____ *1 mark*

(c) Suggest **two** ways Becky could make the tea dissolve more quickly.

1._____

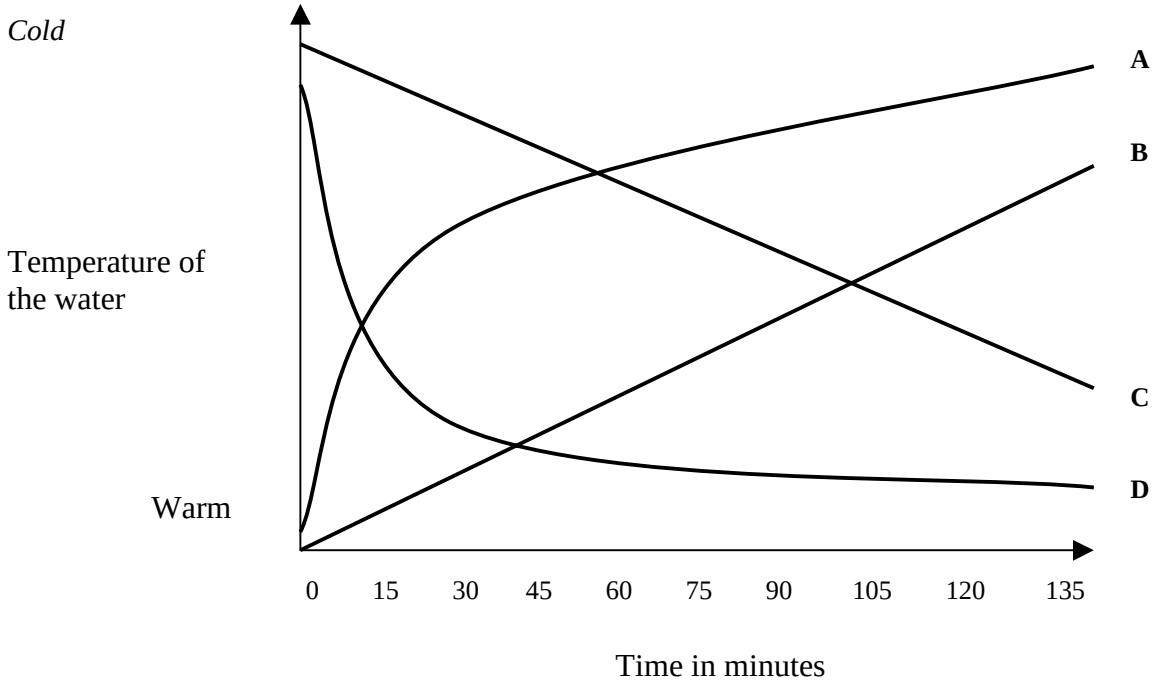
_____

2._____

_____ *2 marks*

(Version 2)

1. In an experiment about cooling liquids Becky poured 50cm³ of warm water into a beaker. She stirred the water slowly. Every 15 minutes she took the temperature of the water.



(a) Which graph, A, B, C or D shows how the temperature of the water changed?

_____ *1 mark*

(b) Suggest why she stirred the water.

_____ *1 mark*

(c) Suggest **two** ways Becky could keep the water warm for longer.

1._____

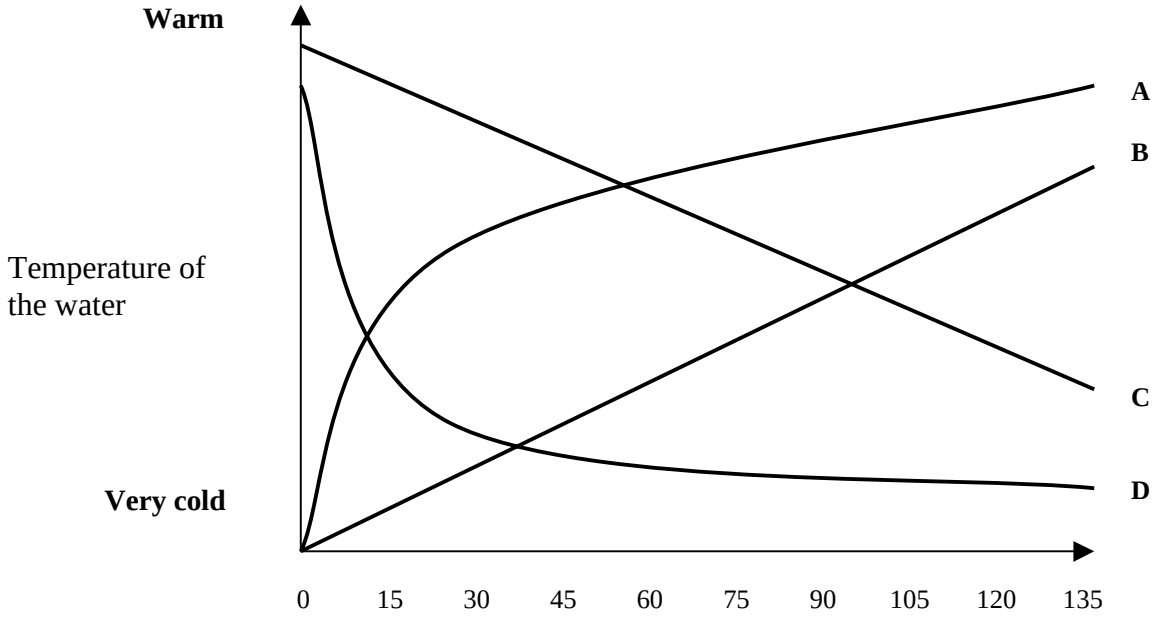
_____

2._____

_____ *2 marks*

(Version 3)

1. In an experiment about warming liquids Becky poured 50cm³ of very cold water into a beaker in a hot room. She stirred the water slowly. Every 15 minutes she took the temperature of the water.



(a) Which graph, A, B, C or D shows how the temperature of the water changed?

_____ *1 mark*

(b) Suggest why she stirred the water.

_____ *1 mark*

(c) Suggest **two** ways Becky could keep the water cold for longer.

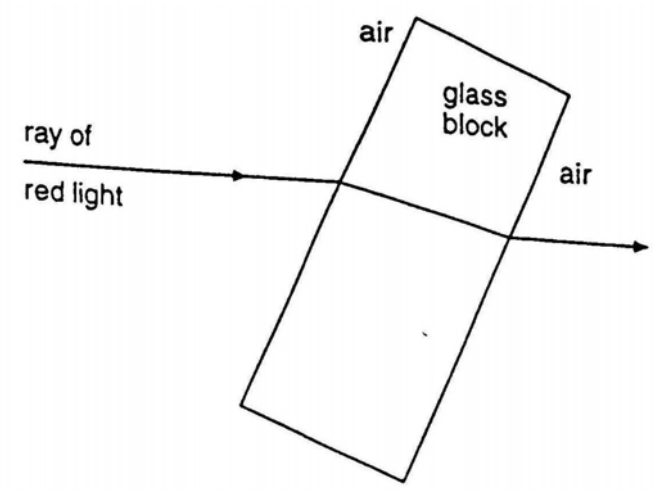
1._____

_____

2._____

_____ *2 marks*

2. (a) The diagram below shows a ray of red light entering a glass block.



(i) Most of the light goes into the glass block, but some does not.
What happens to the light which does **not** go into the glass block?

1 mark

_____

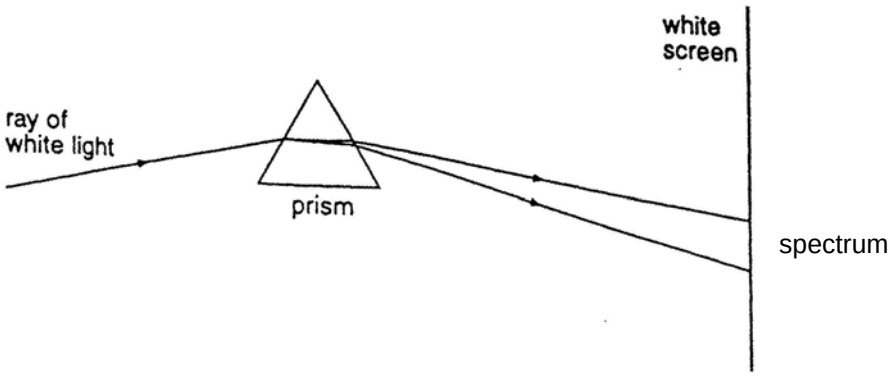
_____

(ii) As the light goes into the glass block, it changes direction.
What is the name of this effect?

1 mark

_____

(b) The diagram below shows white light passing through a prism and forming a spectrum on a white screen.



The spectrum contains light of all colours. Red is at one end of the spectrum. Write **blue, green** and **violet** below in the order of the spectrum.

1 mark

Red

_____

_____

_____

(c) A pupil puts a green filter in the ray of white light. What happens to the spectrum on the screen?

Tick the correct box. *1 mark*

The whole spectrum turns green. ☐

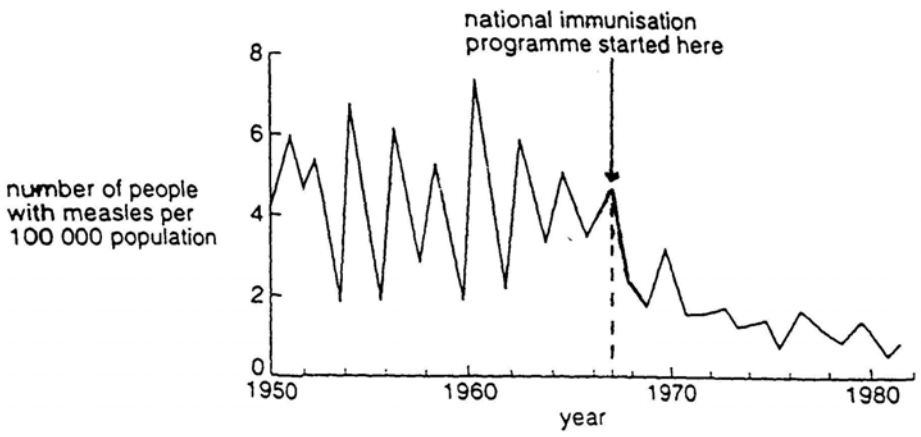
The green part of the spectrum disappears, but the other colours stay the same. ☐

The green part of the spectrum stays the same, but the other colours disappear. ☐

The whole spectrum disappears. ☐

3. (a) In 1967 a national immunisation programme against measles began. Children were injected with a measles vaccine to make them immune to the disease.

The graph shows how the number of people with measles varied between 1950 and 1980.



(Data obtained from *New Scientist*, 18th November 1982)

(i) What does 'immune' mean?

1 mark

_____

_____

(ii) Complete the following sentence. *1 mark*

When a person is vaccinated, white blood cells produce _____ which kill micro-organisms.

(iii) What is present in a vaccine to cause white blood cells to respond in this way?

1 mark

_____

(b) Explain how a new born baby can have immunity for a short time without being vaccinated.

1 mark

_____

_____

(c) The national immunisation programme worked well. Explain how the graph shows this.

1 mark

_____

_____

(d) An increasing number of children are not being immunised. *Predict what is likely to happen to the number of cases of measles as a result of this.*

1 mark

_____

_____

(Version 1)

4. A crate of 12 cans of cola costs £4.20.

How much do 7 crates of cola cost?

_____ *1 mark*

(Version 2)

4. A ski pass costs £4.20 per day.

How much would this cost for 7 days?

.

_____*1 mark*

5. Moles live in underground tunnels which they dig themselves.

They are good at digging, and they eat earthworms and other small animals.

(a) Look at the drawing of a mole. Describe one way the mole is adapted for moving through the soil.

1 mark

_____

_____

(b) (i) Complete the sentence below.

1 mark

Moles use their sense of smell to help them to _____

_____

(ii) Suggest why animals which live underground do not need to have good eyesight.

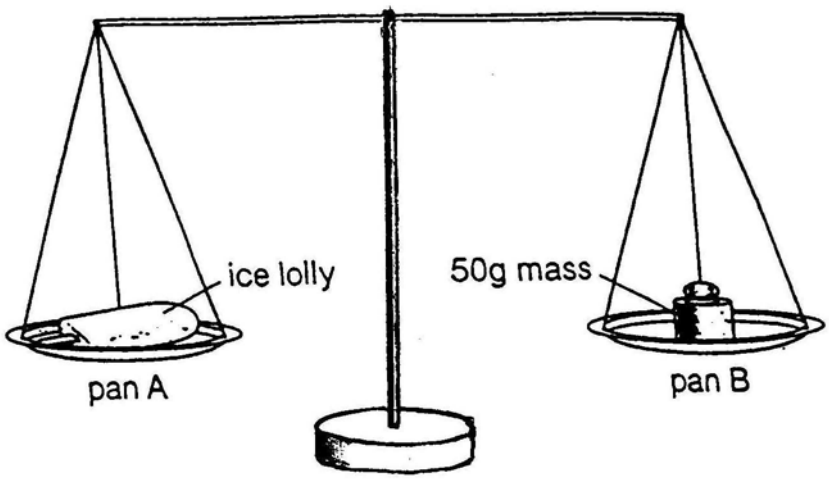
1 mark

_____

_____

(Version 1)

6. Peter put an ice lolly on one pan of a balance.
He balanced this with a 50 g mass on the other pan.
This is shown in the diagram.



(a) Peter left the balance like this for 15 minutes until the ice lolly had melted.
Place a tick in the box that describes the balance after 15 minutes

Pan A has moved up.	☐
Pan A has moved down.	☐
Pan A and pan B are still at the same level.	☐

1 mark

(b) Peter left the balance like this for two days. When he came back, pan A was dry.

(i) How did the position of pan A change?

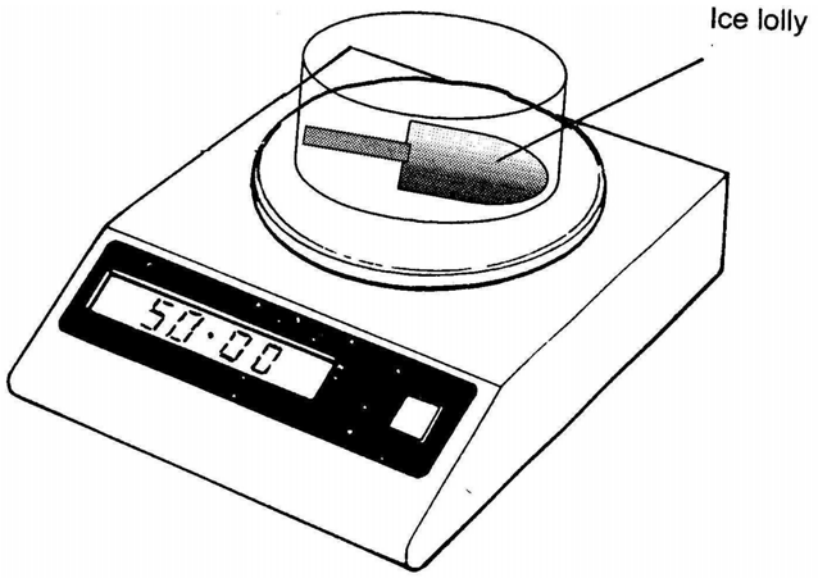
_____ *1 mark*

(ii) What has happened to the water in pan A?

_____ *1 mark*

(Version 2)

6. Peter put an ice lolly on his electric scales.
The scales showed a reading of 50g.
This is shown in the diagram.



(a) Peter left the scales like this for 15 minutes until the ice lolly had melted.
Place a tick in the box that describes the scale reading after 15 minutes

Greater than 50g	☐
Less than 50g	☐
50g	☐

1 mark

(b) Peter left the scales like this for two days.
When he came back the scale pan was dry.

(i) How had the scale reading changed?

_____

1 mark

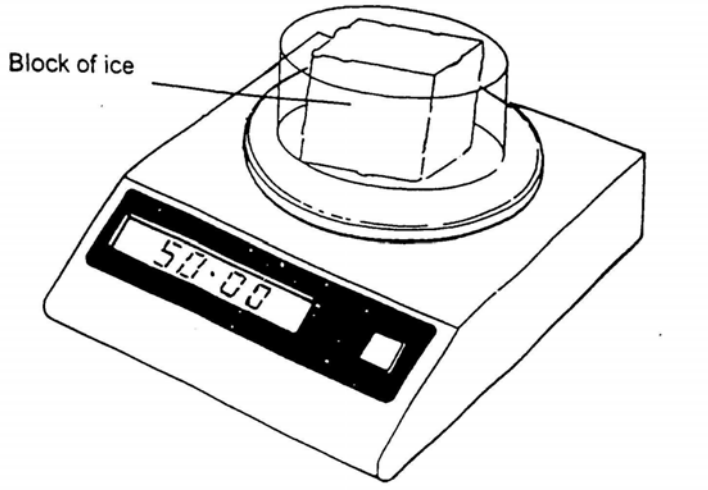
(ii) What has happened to the water in the scale pan?

_____

1 mark

(Version 3)

6. Peter was investigating what happens when ice melts.
He put a block of ice on his electric scales.
The scales showed a reading of 50g.
This is shown in the diagram.



(a) Peter left the scales like this for 15 minutes until the ice had melted.
He then took a reading from the scales.
Place a tick in the box that describes the scale reading after 15 minutes

Greater than 50g	☐
Less than 50g	☐
50g	☐

1 mark

(b) Peter left the scales like this over the weekend.
When he came back the scale pan was dry.

(i) How had the scale reading changed?

_____ *1 mark*

(ii) What has happened to the water in the scale pan?

_____ *1 mark*

(Version 4)

6. (a) A 50g block of ice melts completely.
What mass of water is produced?

More than 50g ☐

Less than 50g ☐

50g ☐

1 mark

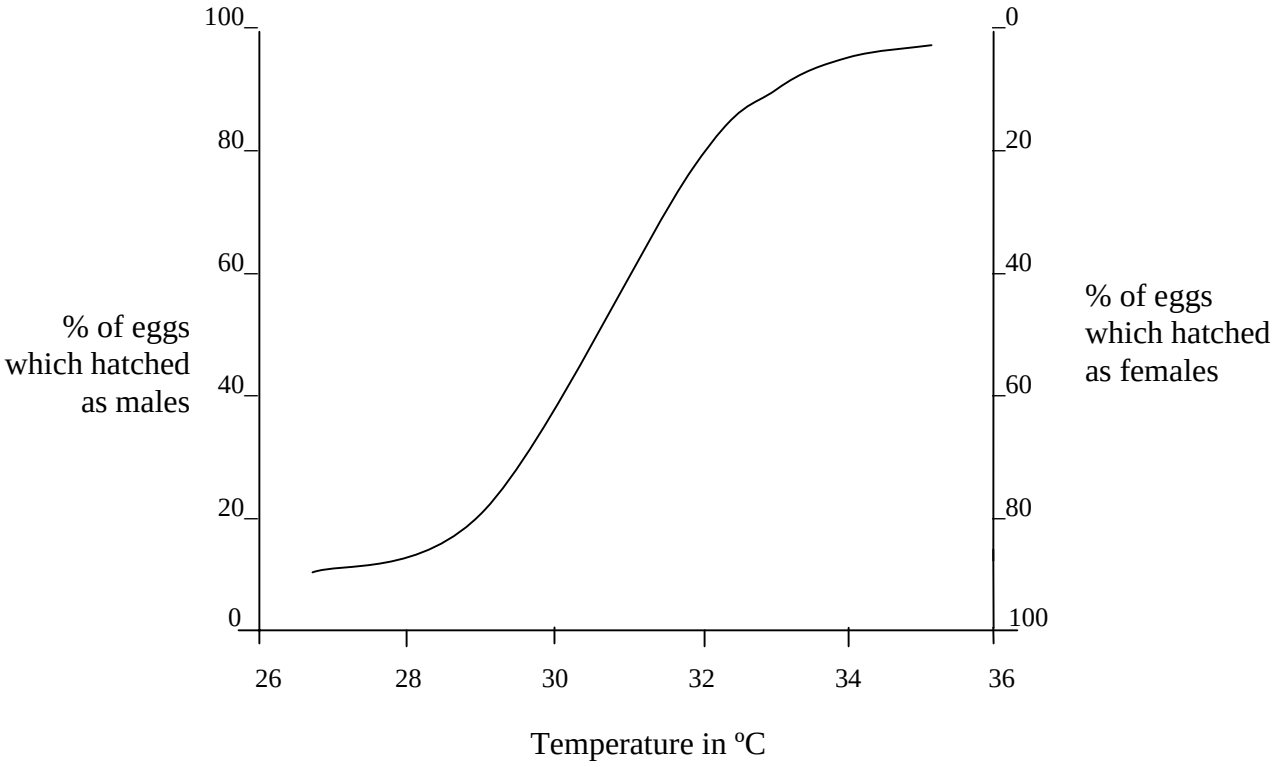
(b) If a dish of water is left uncovered in a room it will dry up.
What has happened to the water?

_____

1 mark

(Version 1)

7. Alligators are reptiles which lay eggs on land.
The graph shows the number of males and females which hatched in an experiment when alligator eggs were kept at different temperatures.



(a) Describe how temperature affects the number of male and female alligators which hatch from the eggs.

_____

_____

_____

2 marks

(b) What percentage of eggs hatch as males at 28ºC ?

_____%

1 mark

(c) At what temperature did 20% of the eggs hatch as females?

_____ºC *1 mark*

(d) How could the zoo keeper make sure that as many females as possible hatch from the eggs?

_____

_____

1 mark

(e) Suggest what would happen to the eggs at a temperature of 50ºC.

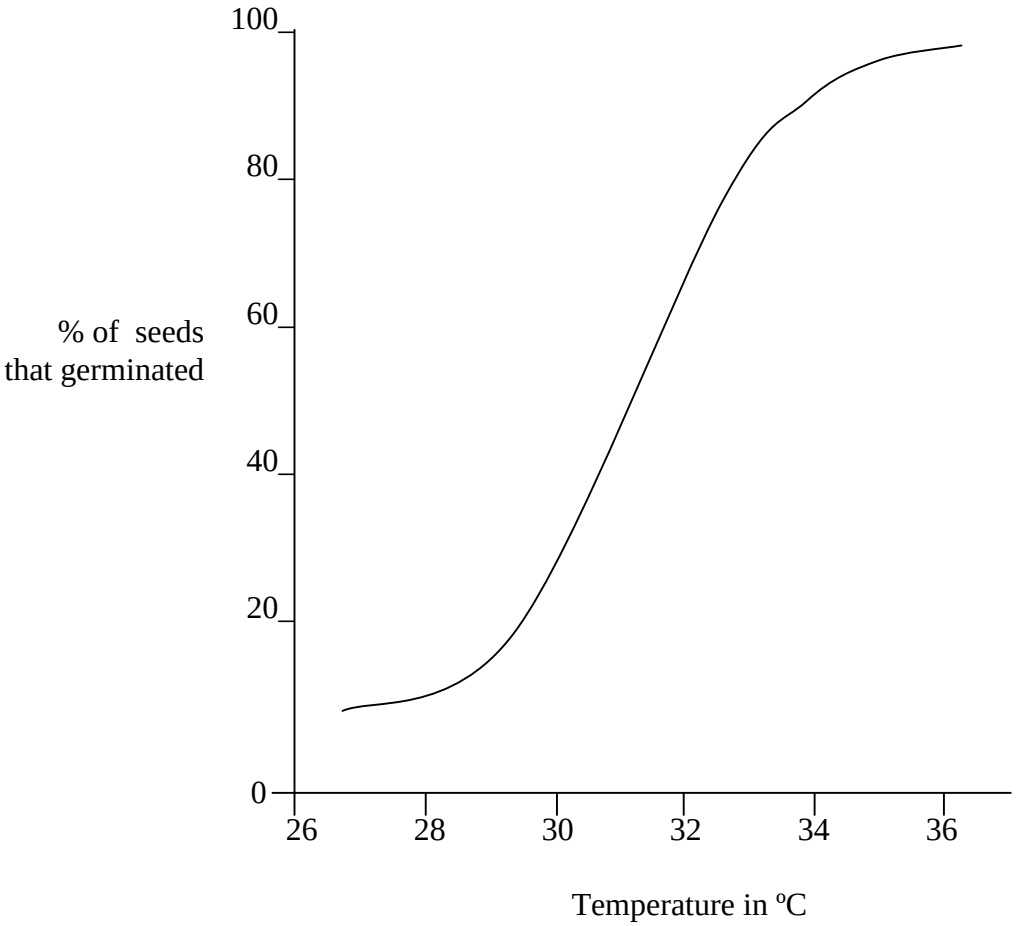
_____

_____

1 mark

(Version 2)

7. Plant growers carefully control the temperature of their seed trays. The graph shows the percentage of seeds that germinate in an experiment when the seed trays were kept at different temperatures.



(a) Describe how temperature affects the number of seeds that germinate.

_____

_____

_____

2 marks

(b) What percentage of seeds germinated at 28ºC?

_____%

1 mark

(c) At what temperature did 20% of the seeds **FAIL** to germinate?

_____ºC *1 mark*

(d) How could the plant growers make sure that as many seeds as possible will germinate?

_____

_____

1 mark

(e) Suggest what would happen to the seeds at a temperature of 50ºC.

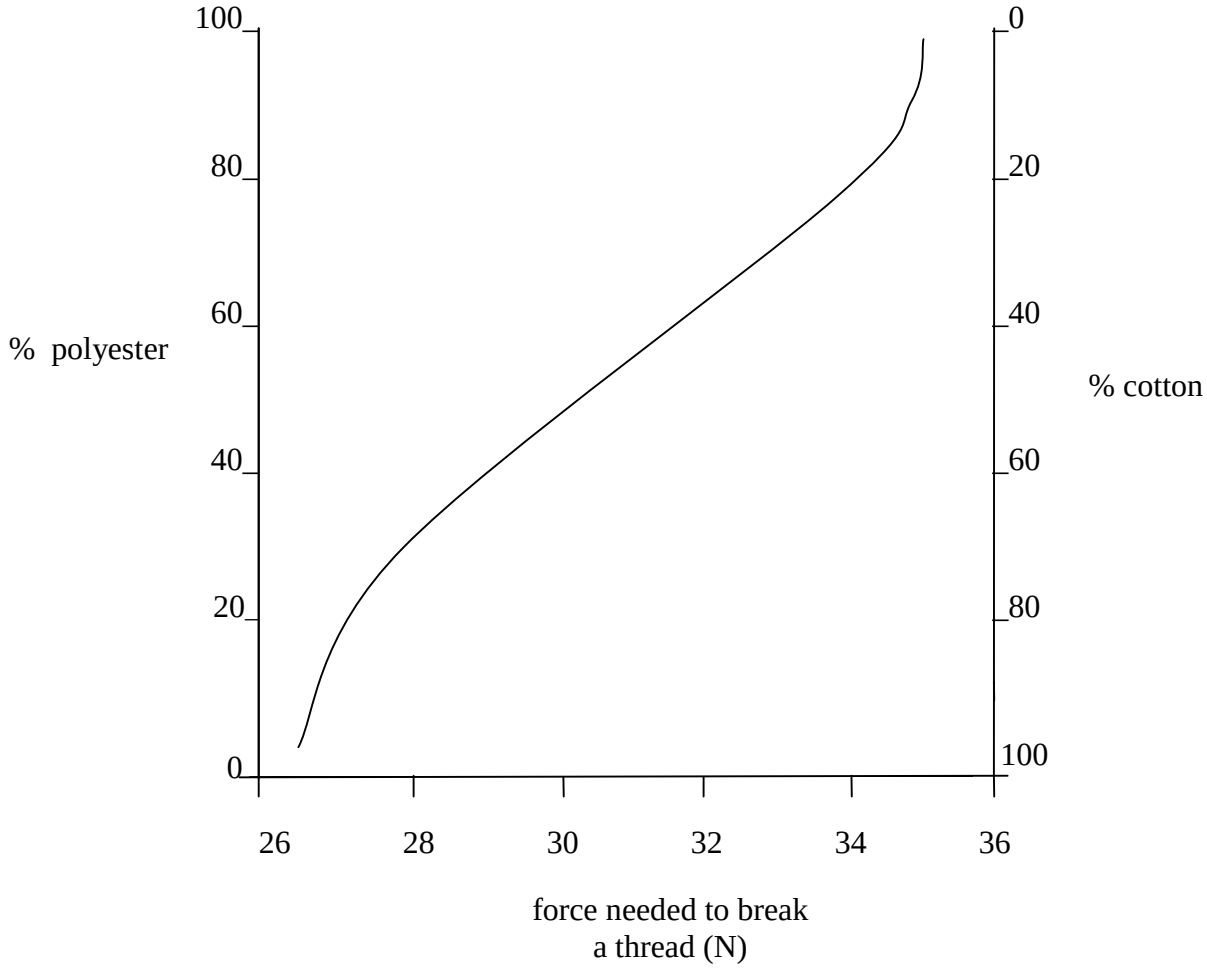
_____

_____

1 mark

(Version 3)

7. Man–made fibres are mixed with natural fibres to create threads for making cloth. The graph shows the force needed to break threads in an experiment with different mixtures of polyester and cotton.



(a) Describe how the force needed to break a thread is affected by the amount of polyester and cotton in the mixture.

_____

_____

_____

2 marks

(b) What is the percentage of polyester in a thread that breaks at 28 N?

_____%

1 mark

c) What force is needed to break a thread containing 20% cotton?

_____N *1 mark*

(d) How could the manufacturers make the threads as strong as possible?

_____

_____

1 mark

(e) Suggest how you could make a thread that would support 50N without breaking.

_____

_____

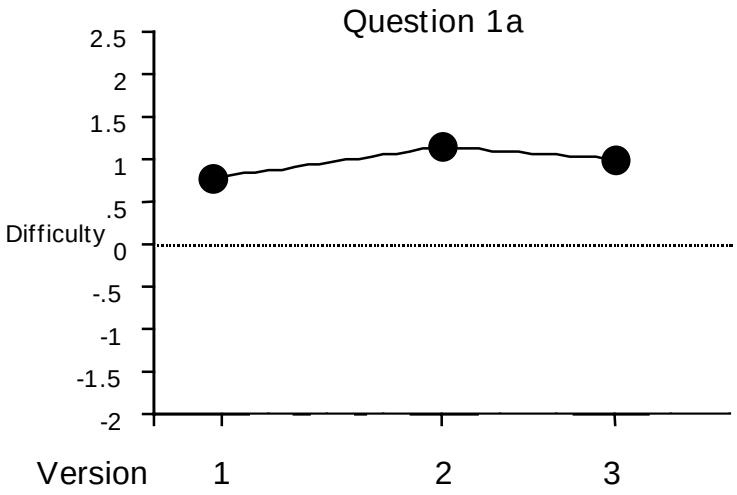
1 mark

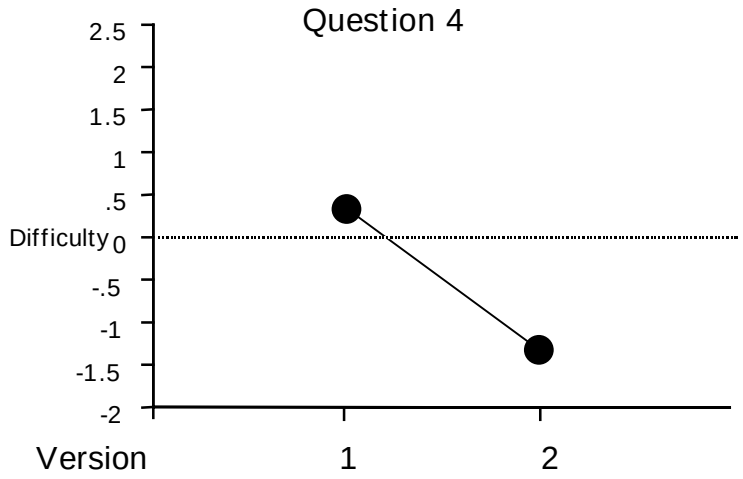
Appendix 2 - the versions

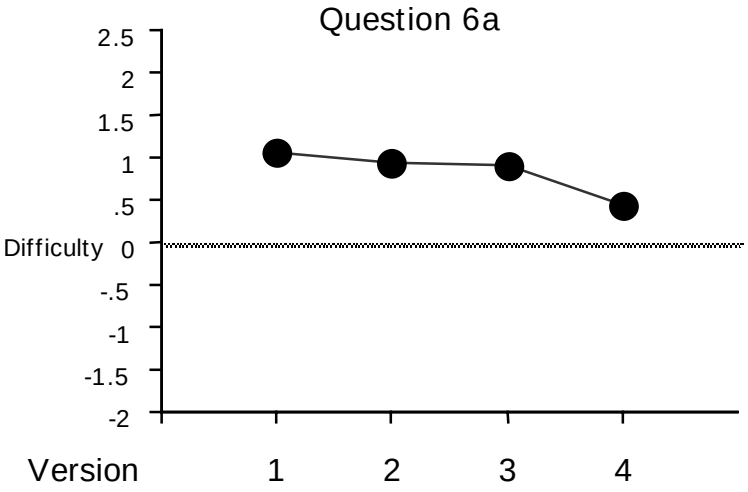
Twelve different papers were made up using the different versions of the questions as follows:

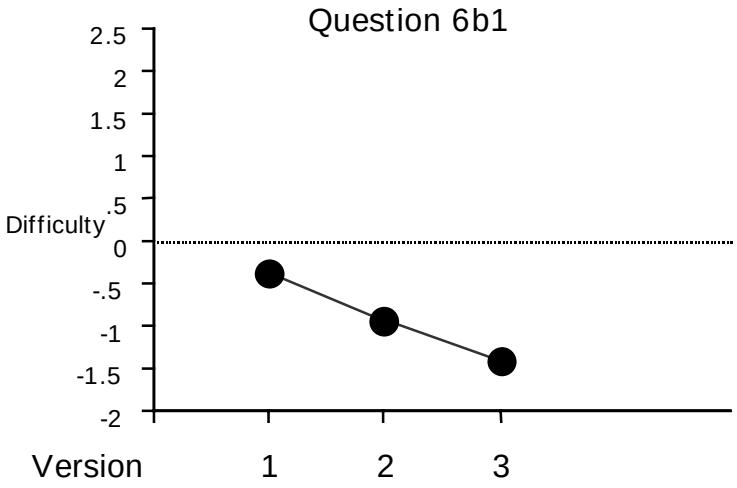
	A	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***	***K***	***L***
Q 1	V2	V3	V2	V1	V3	V1	V2	V3	V1	V2	V3	V1
Q 2	c	c	c	c	c	c	c	c	c	c	c	c
Q3	c	c	c	c	c	c	c	c	c	c	c	c
Q4	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
Q5	c	c	c	c	c	c	c	c	c	c	c	c
Q6	V1	V2	V3	V4	V1	V2	V3	V4	V2	V1	V3	V4
Q7	V2	V3	V1	V2	V3	V1	V2	V3	V2	V1	V3	V1

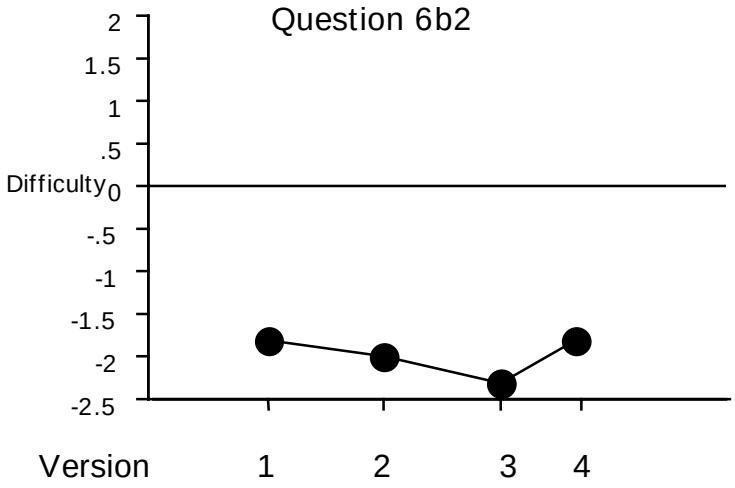
Appendix 3 – Difficulty values in graphical form

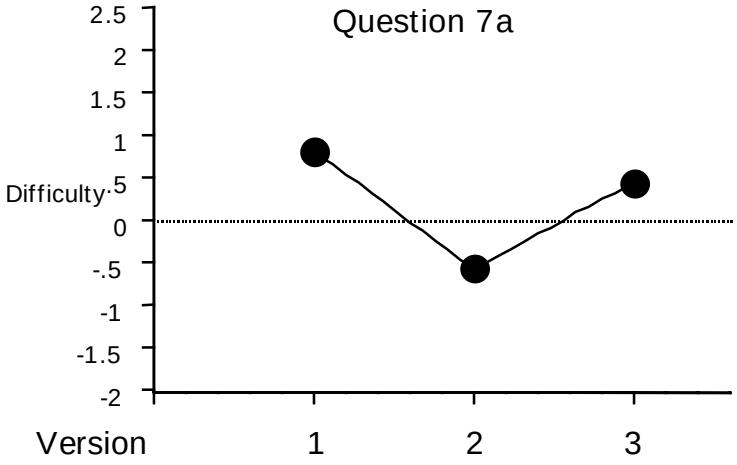


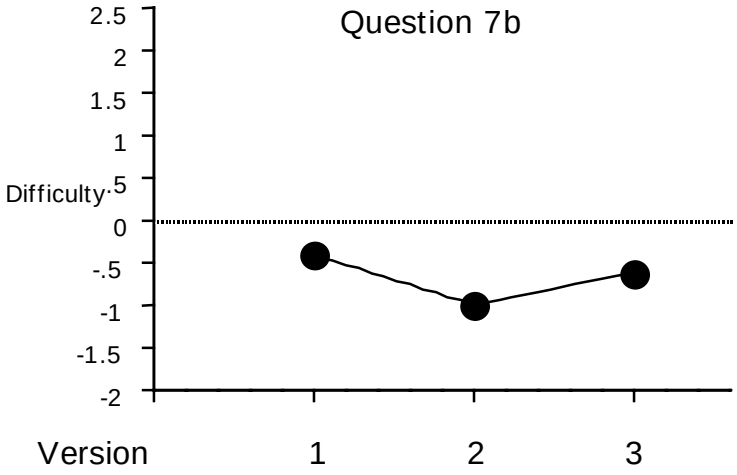


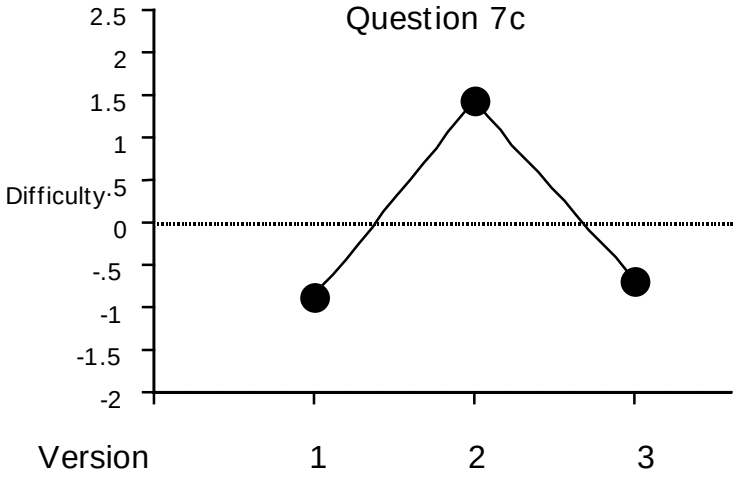


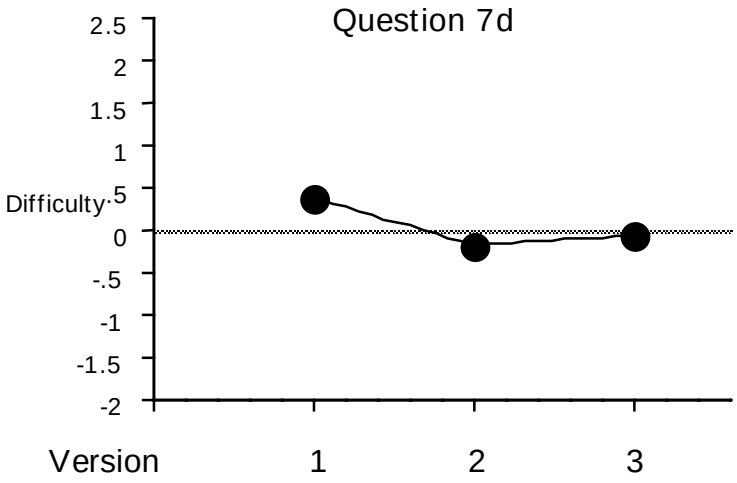

Question 6b1
2.5
2
1.5
1
.5
Difficulty
0
-.5
-1
-1.5
-2
Version
1
2
3

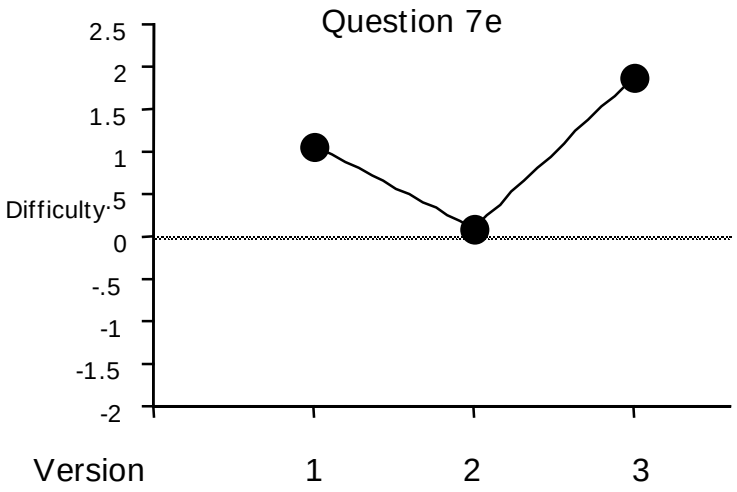

Question 6b2
2
1.5
1
.5
Difficulty
0
-.5
-1
-1.5
-2
-2.5
Version
1
2
3
4


Question 7a
2.5
2
1.5
1
.5
0
-.5
-1
-1.5
-2
Difficulty
Version
1
2
3


Question 7b
2.5
2
1.5
1
.5
0
-.5
-1
-1.5
-2
Difficulty
Version
1
2
3


Question 7c
2.5
2
1.5
1
.5
0
-.5
-1
-1.5
-2
Difficulty
Version
1
2
3





Appendix 4 – Interview questions

The interview questions were as follows:

What are your thoughts about the test?

Was there one question you thought was particularly easy?
Why?
Further prompting if necessary.

Was there a question you remember as being particularly difficult?
Why?
Further prompting if necessary.

Is there any other question you remember well?
W*hy?*

The interviewer then turned to any versions of questions that had not been mentioned and asked the students for their thoughts on these.