Tales of the Expected: The Influence of Students' Expectations on Exam Validity

Ezekiel Sweiry, Victoria Crisp, Ayesha Ahmed & Alastair Pollitt
University of Cambridge Local Examinations Syndicate

A paper presented at the British Educational Research Association Conference, Exeter, September 2002

Disclaimer
The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Contact details
Ezekiel Sweiry and Victoria Crisp
Research and Evaluation Division,
University of Cambridge Local Examinations Syndicate,
1 Hills Road, Cambridge,
CB1 2EU.
☎ 01223 553846/553805
FAX: 01223 552700
Sweiry.e@ucles.org.uk or crisp.v@ucles.org.uk

This paper is available at www.ucles-red.cam.ac.uk

Tales of the Expected: The Influence of Students' Expectations on Exam Validity

Abstract

Through classroom preparation and exposure to past papers, textbooks and practice tests, students develop expectations about an exam: what will be asked, how it will be asked and how they will be judged. Time pressures and anxieties that students face in exams lead to an over-reliance on these expectations. Questions that do not match expectations may result in students getting the wrong answer for the wrong reasons.

Expectations about examinations include the way in which questions will be worded. Students will read what they expect to read. For example, they may overlook negative words (e.g. 'not', 'only', 'a few'). Students will also have expectations about what they are likely to be asked within a subject and the kinds of answers that they will be expected to give. For example, they may have preconceptions about how much they are expected to write.

The influence of expectations means that there is a requirement for teachers to prepare students for examinations in a certain way. But this is problematic in that it encourages 'teaching to the test' and increases the perceived importance of 'exam technique'. In order for our exams to be valid, we should be assessing conceptual knowledge and competence in a particular subject. How can question writers address this threat to validity?

We will illustrate these ideas with examples from past examinations and students' reflections on them.

Introduction

Throughout life our experiences have a big impact on our expectations of what might happen next, from what the cashier will probably say to us when we pay for our shopping to what might happen if we were stranded on a desert island. Our experiences and also the experience of others as gleaned from friends or from the media, prepare us to deal with likely happenings effectively.

When we experience an event or situation initially it is stored as an episodic memory (i.e. relevant just to that specific event) (Conway et al 1997). However, that representation may later be transformed into a generalised 'schema' for that kind of event. Ideas from this schema will then be activated in other similar situations.

Bartlett (1932) described a schema as a pre-defined framework that represents a typical scenario. This framework has been formed as a result of past experiences. We have schemas for events in everyday life, such as eating in a restaurant and sitting an examination. Schemas include 'slots' for the various characteristics of a particular event. For example, in the schema for eating in a restaurant, we may have slots for people, objects, such as waiting staff, a chef, other customers, crockery, menus, and tables. These are the things that we would expect to find in that situation. We therefore do not need to be told all the details, so long as they are consistent with the general schema we all share about restaurants. If somebody was recounting their visit to a restaurant, they would not have to introduce the items mentioned above, in order for the story to make sense, as these are already part of a schema which is familiar to most people. For instance, one would not be surprised to hear them mention '*the* waiter/waitress' (as opposed to '*a* waiter/waitress'). In other words, he or she is not really 'new' because a stereotypic waiter or waitress was already expected. Such schemas are triggered automatically (usually unconsciously) in response to related ideas. This means that it may be harder to react appropriately to circumstances that contradict our expectations.

Students will have developed a number of schemas relating to examinations and examination questions, based on their experiences of classroom tests and past examination papers. These may include a schema of what exams are usually like, more specific schemas relevant to particular subjects and also schemas to do with question styles that may be activated when reading a question. This causes students to expect to be asked to perform in certain ways. They will have ideas about the kind of things they are likely to be asked, the way in which they will be asked, and the kinds of answers that will score marks. Students' expectations usually help them to prepare appropriately for the exam and to cope more efficiently. This can be time saving and allow students to allocate more attention to producing the answer. However, in the same way as in everyday life, a situation that contradicts with an expectation can be problematic and a student may end up answering the question that they expected to be asked rather than the question that the examiner was trying to ask.

Johnson-Laird (1981) describes us as 'cognitive satisfiers', in that once we have constructed a meaning which we think is compatible with a given sentence, we are unlikely to continue searching for an alternative one. This effect is particularly pronounced in an exam, where students are often anxious to 'recognise' the tasks and get on with answering. Written words are processed sequentially and even the first

few words of a sentence or question can trigger further expectations of what will be read next (Frazier & Rayner, 1982). This is a serious factor as it can lead to students mis-reading what is written. If their initial understanding of a question is dominated by their expectations, it is possible that they will not understand the question in the way that was intended, since they are likely to stop generating alternative meanings if they believe this understanding is correct. This can happen even if the question itself is very clearly written.

The more familiar the style or scenario of the question, the more powerful the schema that is evoked by it and the stronger the expectations of what is required. Students' schemas will resemble each other in important respects. They will construct a representation of the task incorporating common expectations of examination requirements. Yet schemas will also vary to some degree between individuals. For example, different students deem different elements of the text to be most salient, which in turn leads to different concepts being activated in different students' minds (especially when the question is set in the context of a real world situation), creating different conceptualisations of the task. This means that different students may end up answering different questions, and that sometimes we never find out whether students could have answered the question that we wanted them to answer. In such a case, examiners have lost control of what they are measuring.

The generation of meaning will be affected by the activation of schemas. It is important to note that schemas are activated automatically and students will not be aware of this. Evans (1989), points out that 'many biases are caused by pre-attentive or preconscious heuristic processes which determine selective encoding of psychologically "relevant" features of the problem'. This occurs before the reading of the text reaches consciousness. In other words, the subsequent interpretation of the text is affected by processes of which we are unaware. On other occasions, if incoming information does not fit with our present schema then the conflicting interpretations which result may come to consciousness in order that a meaning can be reached.

We can also think about expectations in terms of stereotypes. While schemas are created to cover ideas about events and scenarios, stereotypes reflect features/characteristics of objects or concepts. Through experience, our minds build stereotypes (or prototypes, Rosch, 1978) of certain things, the most well known of which are social stereotypes of people. For example, we all have our own stereotype of what a policeman will be like, perhaps: tall, smartly-dressed, well spoken, and reliable. In the same way students have stereotypic ideas of what exams are like or what a question on a particular topic will be like. These stereotypes are cognitively necessary to allow us to deal efficiently with vast amounts of incoming information. Often these built in stereotypes can be beneficial as they help students to understand quickly what the question is about. However, sometimes these stereotypes, like stereotypes of people, can be wrong and may cause students to do the wrong thing when their stereotypical expectation of something is contradicted.

Ahmed and Pollitt (2001) have discussed construct validity in examination questions. They argue that ‘a question can only be valid if the students’ minds are doing the things that we want them to show us they can do’. In other words, validity requires that students get the question right or wrong for the right reasons. As already

mentioned, students' expectations can lead them to an interpretation of the question that was not intended by the question writer. In such cases, question validity has been compromised, as we never find out whether the students who answered a misinterpretation of the question could have answered the question that was intended.

Particular types of questions and features of questions are more likely to trigger students' schemas and hence expectations. In some cases these expectations may be detrimental to their conceptualisation of the task requirements. In order for our exams to be valid, we should be assessing conceptual knowledge and competence in a particular subject. Other factors should not affect performance. This study sets out to investigate whether we can control the effects of students' expectations by altering question style and wording, with the aim of reducing the threat to validity.

We will first describe the different levels at which expectations in exams seem to take effect, using example questions to illustrate. We will then move on to describe our current investigation of these issues.

Levels of Expectation

Extensive analysis of examination scripts at all levels has led us to believe that students' expectations can influence their understanding of questions at a number of levels: the level of the subject, the question and the sentence.

1. Subject Level Expectations

Students will have expectations of the kinds of things that will be asked within a subject, or even within a topic. If a question includes content that is not expected within a particular subject area, then students may try to interpret the question in a way that they think is relevant to the subject, but one that is not what the examiner intended.

The question below appeared on a Science GCSE paper.

> Suggest a reason why the inventors of this system decided to use a mixture of copper and silver for the electrodes. [1]

The correct answer was that silver is too expensive. However, students did not expect to have to answer in terms of costs in a science paper. Only 10%[1] gained credit on this question. Many tried to answer in terms of the physical and chemical properties of the metals. Also, the question did not ask specifically why using a mixture was desirable rather than a single element, so some students answered with reference to both metals e.g. 'because they both conduct electricity.'

[1] This data comes from a previous internal study where original exam questions and manipulated versions of them were retrialled in schools (Bramley et al., 1998)

A manipulated version of the question, shown below, made the question requirements more explicit, and the number of correct responses increased to 30%[2]. Nevertheless, many students still tried to answer in terms of physical properties as they did not expect an economical argument to be correct in a science exam.

> Suggest a reason why the inventors of this system decided to use a mixture of copper and silver for the electrodes, rather than silver on its own. [1]

2. Question Level Expectations

Students may also have expectations regarding specific features of questions. We have observed several different features of questions that seem to be affected by expectations. These are discussed below.

Space Allocated

The amount of space allocated for a response influences students' expectations about the type of answer required, and the amount that they need to write in order to gain credit.

The example below appeared in a Design and Technology (Automotive Engineering) GCSE paper.

> Complete the table below by naming a material or giving a reason. The first one has been done for you.

Part	Material	Reason why this material is used
Engine mounting	Rubber	absorbs the vibration from the engine
Engine block	Aluminium	
Electric wiring		good conductor of electricity
Radiator		good conductor of heat
Windscreen		will not shatter when broken
Dashboard	Plastic	

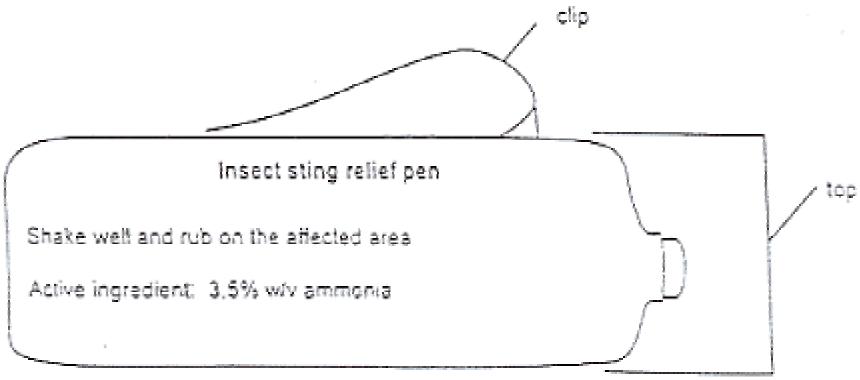
The mark scheme answer for choice of windscreen material was 'toughened/ laminated glass'. A large number of students just wrote 'glass'. In some cases, this may have been because the space allocated for the answer does not appear to be wide enough to fit in a longer answer hence students expected only a short, one-word answer was required.

Relevance of Resources

[2] See note 1.

A common assumption that students hold is that information from question resources (e.g. diagrams, graphs, additional text) should be used in their answer. This often leads students to place more emphasis on resources than was intended by the examiner.

The example below is taken from a science GCSE paper:

> Joe's parents bought an insect relief pen.
>
> 
>
>
> When the top was removed the contents could be smelt from a distance. Use ideas about particles to explain this. [3]

Information relating to shaking the pen and the properties of ammonia were irrelevant to the requirements of the question. Some students used this salient but irrelevant material in their answers and hence lost marks. Some talked about particles bursting out of the tube because it had been shaken, perhaps thinking it was similar to shaking a can of fizzy drink. Others said the particles could be smelt from a distance because ammonia has a bad smell, picking up on the 3.5% ammonia given in the context. Question writers need to be aware that using diagrams and pictures just to illustrate a question can lead students to use these in their answers.

Level of Difficulty

Students also have expectations regarding the intended difficulty level of the question. This may be based on aspects such as the level of the paper (e.g. foundation vs higher, GCSE vs A-Level), the position of the question (sequentially) within the paper, and experience of similar papers in the past.
These expectations influence the level of answer a student will give and are usually helpful. However, sometimes a student's expectations may lead them to 'pitch' their answer at a level that is higher or lower than the one intended by the examiner

The question below is taken from a Sociology GCSE paper.

> 'A pressure group consists of a group of people who have a common interest or concern. Some groups try to promote a cause

whilst others try to protect the interests of their members. The Automobile Association and Shelter are examples of pressure groups'.

Study the source above.

1 (a) (i) Name one pressure group.

(ii) Identify two types of pressure groups.

The demands of part (i) are very low: students are simply required to name one of the two pressure groups named in the preceding text, and normally the vast majority would easily be able to score the mark. However many students, possibly not expecting a question with such low demands, failed to score the mark. Some left the question out altogether, possibly not understanding what was required of them, while others tried to name a group from their own knowledge (not mentioned in the text), but were unable to name a valid one. Part (ii) followed a similar pattern. Students were simply required to re-state the second sentence in the text - that ‘some groups try to promote a cause whilst others try to protect the interests of their members’. Nevertheless, many failed to realise they were simply required to locate the answer from the text, and attempted to come up with an answer from their own knowledge, such as 'direct and indirect groups'.

Mark Allocation

Students tend to interpret marks as a point for a point, when this is not always the case.

The question below is taken from a Geography GCSE paper:

What is meant by the term 'rural depopulation'? [1]

The mark scheme demanded that students gave another word for *both* 'rural' and 'depopulation' in order to get the mark. But the allocation of only one mark to this question implies that only one of these words needs to be explained. Most did not think it was necessary to give another word for 'rural'. If two marks had been allocated candidates may have realised what was required.

3. Sentence Level Expectations

Students’ expectations regarding what will be read next may be influenced by preceding words or sentences in a question. They may also be influenced by past experiences of questions. For example, students may be influenced by the fact that most questions are written in a positive rather than negative form.

The question below is taken from a Mathematics GCSE paper:

On a production line at the SMP chocolate factory, the chocolates are made by two machines. The first moulds the centre and the second coats the centre with chocolate.
Dipti works for the quality control section. Some of the chocolates are faulty. She estimates that the first machine makes a misshapen centre with the probability 0.05 and that the second machine fails to coat a centre with the probability 0.02.

(a) What is the probability that a chocolate picked at random
  (i) is coated, [1]
  (ii) is not faulty [2]

On part (a)ii students tended to overlook 'not' and to calculate the probability that the chocolate was faulty. The question was changed by emboldening the word 'not'. The percentage of correct answers increased from 7.6% to 31.1%[3].

We will now describe a study in which aspects of science questions were manipulated in order to investigate the effects of students' expectations on their performance.

Method

Participants

198 students (103 boys and 95 girls) aged 15 years completed either of two versions of a test paper. They were all studying intermediate level science at local comprehensive schools. Ninety-eight students sat version 1 of the test and 100 students sat version 2 of the test. The two versions were assigned randomly.

The students expected grades at GCSE ranged from A* to F, with 89.4% having been predicted B, C or D. Fifty-two students were interviewed in pairs immediately after they had sat the test.

The questions and question papers

Six questions were selected and slightly adapted from past foundation tier (grades G-D) GCSE (16+ exam) papers. Questions were chosen to be appropriate for students of the chosen age group, avoiding content they would not have covered. We chose questions where we thought students' expectations might have affected how they answered and led them to make mistakes. Four questions were from combined Science papers and two were from Health Studies papers. The Health Studies questions were included because they were interesting in terms of expectations and we could not test these with Health Studies students, as the entry levels for this syllabus are low. They are based around issues covered in biology and therefore it was appropriate to use them in a science paper.

[3] This data comes from a previous internal study where original exam questions and manipulated versions of them were retrialled in schools (Hughes, S., Fisher-Hoch, H. & Pollitt, A. (1998a)

The original questions were manipulated by rewording or changing some of the question parts in a way that we hoped would reduce the students' likelihood of being led astray by their expectations. Comparing performance between the two versions would allow us to see if our adjustments did indeed reduce the impact of expectations.

The six questions, each with 2 different versions were compiled into 2 different versions of the test paper, each containing half of the original versions of questions and half of the manipulated versions. A total of 25 marks were available on each of the versions of the paper.

We decided not to counterbalance the versions because full counterbalancing would require an impractical number of versions. Instead we aimed to distribute the two versions randomly within each class of students and hence make the groups doing each version as equivalent as possible.

Procedure

The tests were completed in the students' normal science classroom or laboratory during lesson time and were carried out under examination conditions. In all cases students were not aware that they were going to be involved in this research before they arrived at their lesson. In one way this is a limitation to our research since the students will not have revised, as they would have for an exam. They will not have gone through preparation for an exam with the teacher or practised lots of test questions recently, all of which are likely to contribute to the expectations that students take with them into an exam. However an authentic preparation would be hard to achieve and would be much more demanding of schools' time. Students who are actually about to sit their exams will probably have much more developed expectations than the students that were involved in our study due to the amount of practice tests and preparation they go through in the months preceding examination.

For ethical reasons we did not want the test to be too stressful although we did want students to try their best. Students were told that we were testing out the questions and that we were carrying out research into how the way that questions are worded affects how students perform. We did not mention our interest in their expectations of exam questions and of exams in general as this might have led the students to act unnaturally. It was made clear that the research was being conducted to help the researchers find out more about asking exam questions and that the aim was not to test them, the students. They were also told that the marks would not go towards their school grades but that the marks would be returned to the school. The latter was an attempt to ensure that students did answer sensibly. We also told the students that there were different versions of the papers with slightly different questions. This was both to be honest with them and to discourage copying. Students were asked to answer all the questions in the test and were given 25 minutes to do so.

Pairs of students were interviewed immediately after the test. The students were either chosen by the teacher or asked to volunteer. The only suggestion that we made to teachers on the choice of students was that they should be comfortable in talking to a visitor. The interviews were carried out in a quiet room away from the classroom.

Interviewees were given their papers to have in front of them whilst they talked. The interviews were tape recorded and then later transcribed.

The purpose of the interviews was to try to find out what the students thought of the questions, their general reaction to them, and to try to find out how the students went about answering the questions. We were interested in what may have led them to make certain mistakes and what made them think they should answer in a certain way. We also wanted to find out if what the students generally expect from a test conflicted or aided them with these questions. Students were interviewed in pairs rather than individually, both to allow us the opportunity to talk to more students and to make the experience less stressful for them. We have found in previous similar research situations that the students prompt each other and this helps to elicit more comments from them (e.g. Ahmed and Pollitt 2001). The interviews were semi-structured in nature with the intention of some qualitative analysis of comments. Three different interviewers took part. The interviews were based around a common framework (see Appendix 2) but were not strictly standardised as this did not seem necessary given the nature of the analysis of the interviews. The comments that students made about the questions will be discussed later in this paper.

Ability of the students

We obtained estimated Science grades for the students from the schools. We then turned these grades (A*-F) into a score where the difference between 2 consecutive grades is 2 points. Where a student was allocated for example a C/D they were allocated a score between the scores for C and D i.e. 5.

Grade	A*	A	B	C	D	E	F
Score	12	10	8	6	4	2	0

We used these measures to find the mean science ability of the students who sat each version of the paper. The results are shown in the table below.

Est. grade.

Version	Mean	N	Standard Deviation
1	5.64	98	1.890
2	5.55	100	1.992
Total	5.60	198	1.938

As the table shows, the mean ability score was very similar for each of the groups.

Analysis of data

The students' papers were marked according to the original mark schemes for each question. A cross-tabulation analysis was run to see the effects of version on scores on each question part. We also coded students' answers by the kinds of responses that were given and ran cross-tabulations on these data to see if the differences between two versions of the same question affected the kinds of answers that students gave. The results of these analyses will be quoted when discussing the questions.

We also ran a Rasch analysis on the students' marks. The results confirmed those of the cross-tabulations mentioned above.

The interviews were not formally analysed but we looked for common comments on questions and will report some examples of these when we discuss the questions in the following section.

Results and Discussion of Questions

(The questions and mark scheme can be found in appendix 1)

Expectations relating to change of topic (Question 4)

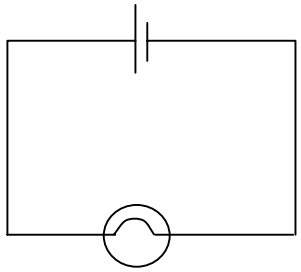
In question 4 we were predominantly interested in part (c).

> (c) One day Joy finds that even with the right answer, the bulb does not light.
> Suggest **two** reasons why the bulb does not light.
> 1 _____
> 2 _____[2]

This question was included in this study with the intention of testing the effect of earlier question parts on latter ones. The question, taken from a combined science paper, began by testing knowledge of biological concepts to do with parts of the body. This was done in the context of someone playing an electrical game, and circuit diagrams were provided. The last part (shown above) was included by the question writer purely to test physics, and required no use of earlier information. It is thought that students would expect there to be a clear link between all parts of a single question, particularly when the context remains unchanged.

This question part was manipulated to try to make it clearer that initial context was irrelevant to this question part; this question was purely testing knowledge of circuits. Part (c) was changed as shown below.

> (c) Look at the circuit diagram below.



> The bulb usually lights but one day it does **not** light.

Suggest two reasons why the bulb does **not** light.
1_____
2_____ [2]

There were three possible marking points for part (c): – 'bulb broken', 'battery flat' and 'wire broken / circuit incomplete'. The likelihood of scoring a mark on the first two of these increased significantly in the manipulated version (29.0% compared to 76.5% and 31.0% compared to 57.1%). Performance on the third marking point was fairly similar on both versions (38.0% compared to 35.7%). Some students were obviously confused by the context of the previous parts. 22.0% of responses in the original version of the question referred to biological issues or to the game context. E.g. "there is a blood clot", "body part cannot carry electricity", "blood or gases won't light a bulb". Their confusion was also evident in some of the interviews, where one student said "That one was a bit confusing because it is mixing biology with physics, so I wasn't sure which one to answer, what is the reason, is it to do with the human body or the circuit?" The manipulated version definitely encouraged students to write answers in terms of physics (number of physics related responses increased from 59.5% to 95.4%).

Sentence Level Expectations
(Question 6)

Part (b), shown below, was of particular interest in this question.

(b) What percentage of injured German athletes did not recover within 4 months of breaking a leg?
…… [1]

This is part of a question adapted from a Health Studies GCSE paper. When this question appeared in a live examination the word 'not' in the question was frequently missed, so that a number of candidates answered the question 'what percentage of German athletes *did* recover within 4 months of breaking a leg'. This may be because of students' expectations that questions will usually be written in the positive form. The manipulated version of this question was identical except that the word 'not' was emboldened with the aim of reducing this problem. The percentage of correct answers increased from 29.0% to 34.7%. This indicates that emboldening 'not' did help to some degree but this improvement is quite small in comparison to other questions that have been manipulated in a similar way (See earlier example in the 'Levels of expectations' section). It is hypothesised that the small difference in this particular case was due to the large number of concepts that needed to be considered in answering this question, meaning that students had to read the question carefully anyway.

The students' responses were analysed to measure the proportion of students who spotted the word 'not'. Even in the manipulated version where the word 'not' was emboldened, over a quarter of students (26.5%) still overlooked the word 'not'. This indicates that students' expectations were still threatening the validity of this question despite the change that we made.

Subject and difficulty related expectations
(Question 5)

Read the following passage.

Rachel is 15 and gets upset very easily. She is always arguing with her mother and older sister. She spends a long time in the bathroom or in her room trying out new hair styles and make-up. Rachel likes to go shopping with her friends and spends most of her pocket money on CDs. She also buys cigarettes and hides them at home. She has just bleached part of her hair. Her mother buys most of her clothes which are not fashionable enough for Rachel. When she goes out Rachel often borrows her older sister's clothes and make-up without asking and changes in the ladies' toilets. Rachel and her friends spend a lot of time standing chatting to older boys and smoking in the shopping centre near the main music store.

(a) Use the information in the passage to suggest **one** cause of Rachel's arguments with her mother or sister.

_____
_____[1]

(b) Suggest **one** way Rachel's mother could help to reduce the number of arguments.

_____
_____ [1]

This question came from a Health Studies GCSE paper. Students often seemed to expect answers to be based on specific subject knowledge. In this case, common answers to (a) included 'puberty - her ovaries are producing hormones that can affect her behaviour' and 'she needed an intake of nicotine'. In fact, the mark scheme answers required simple comprehension of the passage e.g. 'Occupying bathroom', 'Borrowing clothes without asking'. It is possible that students did not expect to gain marks for comprehension type answers in such a question. One student said "I think that one was easy. I think I got a bit, well I didn't get confused but I thought that that wouldn't be in a science paper, it would be in an English paper." Another commented "I was going to put my own knowledge, like about hormones and growing up and the body, but it said use information from the passage. I didn't really understand it very well." The fact that a large number of students noticed that the question was not a typical science question demonstrates students' expectations regarding the scope of questions that will figure in a particular subject.

It is also possible that simple retrieval of text seemed like too easy a task to be worthy of marks on a GCSE paper. This may have been part of the reason why some students attempted the more difficult task of inferring an underlying cause for Rachel's behaviour.

The question was manipulated by exchanging the word 'suggest', which was thought to prompt students to use their own ideas, with the word 'state'. The aim of this was to help focus students on ideas that were already in the text.

Performance increased on part (a) from 53.1% to 61.0% of students getting the mark. This difference suggests that the word change was partially successful in clarifying the question requirements. Performance on part (b) increased slightly from 72.4% to 75.0%.

Subject related expectations
(Question 2)

> A factory makes ammonium nitrate.
> Write down **three** costs of making ammonium nitrate in the factory.
> 1. _____
> 2. _____
> 3. _____[3]

In question 2 we were particularly interested in part (b)ii (shown above). The term 'costs' can have a variety of meanings. In this question the answers required by the mark scheme were economic costs such as 'raw materials', 'wages', 'energy/electricity' and 'plant costs'. This probably contradicts most students' expectations at the subject level, or in other words students are unlikely to expect to be able to gain marks for writing about economic factors in a science paper. Our manipulated version of this question was identical except that the word 'costs' was replaced with the word 'expenses'. The aim of this was to encourage students to focus on economic ideas.

The change brought about no real difference in performance. The word 'expenses' did not seem to help students understand the kind of answers they were expected to give. In both cases some students seemed to be confused as to what the question really meant. Some students wrote about the conditions that are required for the reaction to take place (e.g. high pressure and temperature), some answered in terms of environmental costs (e.g. harmful to the earth's atmosphere) and others took costs to mean disadvantages of the process (e.g. it's dangerous). For example, one interviewee remarked that he had "just put down like yeah like its very expensive because like they poison the atmosphere, there are costs to the atmosphere". We do not know whether these students knew the correct answer or not.

Nearly a third of students left this question out (31.0% in the original version and 30.6% in the manipulated version). When interviewing students it became apparent that some of them who had left it out were thinking about economic ideas but didn't write anything because they didn't expect to get marks for such responses.

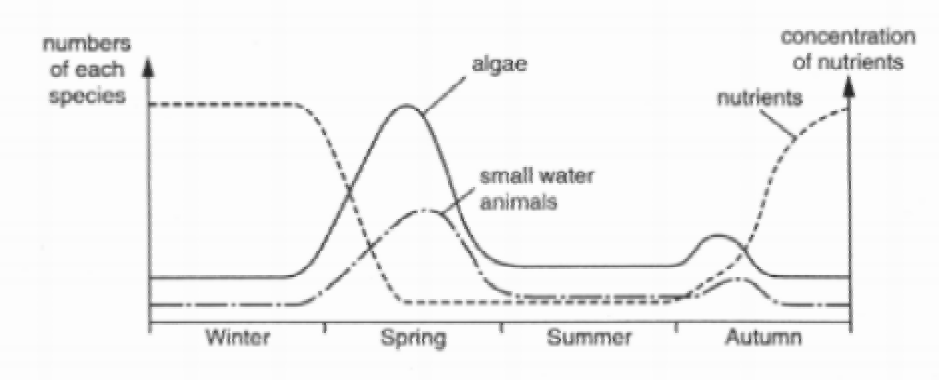
Our wording change did not help allay students' expectations. This was obvious during interviews. One student said "I found that quite confusing since it's not ideal, it's not as you expect, expenses it sounds more like economics or managing money, it's not science really". It is difficult to know whether the question could have been manipulated successfully in order to focus students' ideas on economic costs, due to the strength of the stereotype that students may hold that science and economic considerations do not relate.

It is a requirement in some Science syllabuses for economic costs to be considered, and this is obviously why this question appeared. However, it seems that it is very difficult for examiners to assess this area when students cannot suppress their expectations and understand what they are really being asked. The problem is particularly pronounced in this example because students who don't realise what to do will lose three marks. One potential way to reduce the problem would be to include an example answer and then ask for students to provide two more examples. This might have improved the validity by confirming to students who were unsure that monetary answers were necessary.

Expectations of the relevance of resources
(Question 1)

This question is about conditions in a garden pond.

The graph shows the levels of green algae, small water animals and soluble nutrients in a garden pond over a period of one year. The small animals feed on the green algae.



(a) When is the concentration of nutrients in the pond highest?
..... [1]

(b) (i) Suggest two reasons for the rapid rise in the level of green algae in early spring.
1.....
2 ..... [2]

(ii) Suggest why the level of green algae falls rapidly in late spring.
.....[1]

Part (a) and part (b)ii of the original question required the student to use information from the graph in order to answer the question, while part (b)i required them to use their knowledge. The correct answers to part (b)i were 'increasing temperature / warmth' and 'increasing light / more daylight / more light'.

We believe that students' expectations regarding the *relevance of resources* may have led them to expect to have to use information from the graph in order to answer part (b)i, and they would therefore have lost marks. Students' over-reliance on the graph may have also been exacerbated by the ordering of the question parts; the question parts immediately preceding and following part (b)i did require students to use the graph.

In our manipulation of the question, we tried to counteract this problem by adjusting the question order (parts (b)i and part (b)ii were swapped round and renumbered).We also altered the question wording of part (b)i, by beginning with the phrase 'Using your **own knowledge**'.

These changes had no significant effect on performance on part (b)i. Notably, when students knew that they needed to use their own knowledge they were more likely to leave it blank. The percentage of students leaving the question out increased from 1.0% to 16.0%.

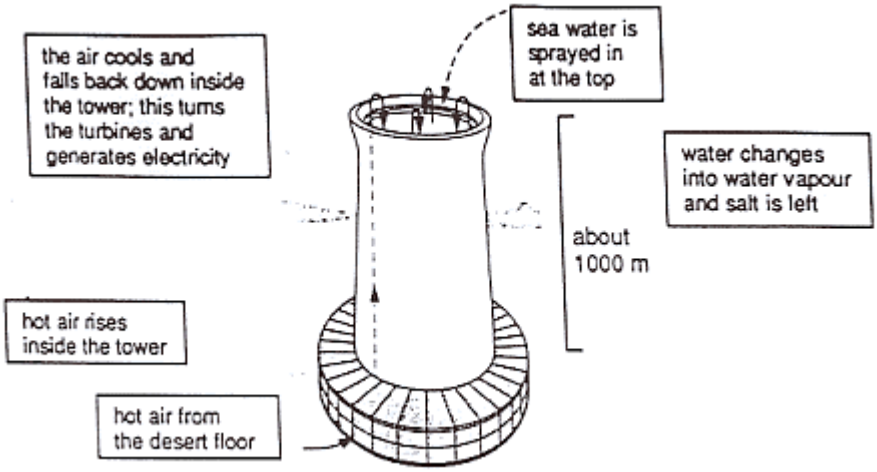
We were surprised to find that our changes had not encouraged more students to use their own knowledge. 47.4% of responses in the original version and 46.5% of responses in the manipulated version used own knowledge. However, our changes did reduce the number of responses taken from the graph from 42.4% to 30.0%. This may be because students did not know what kind of knowledge was intended. They might have thought that they needed to use some learnt scientific knowledge, not just that it gets warmer and lighter in spring.

In part (b)ii (part (a)ii in the manipulated version) the number of students scoring the mark increased in the manipulated version where the question part was closer to the graph. The percentage of students scoring the mark increased from 42.9% to 48.0%. More students seemed to realise that they needed to use the graph in their answers. In further analysis of individual responses it was found that more students used the graph when answering this question if they had the manipulated version (70.0%) than if they had the original version (63.3%). However, these figures are not significant.

Expectations of the relevance of resources
(Question 3)

The original question (below) contained a complicated context but the students did not need to use this in their answers. This may have conflicted with students' expectations about the *relevance of resources*. The required answer to part (a) was a 'textbook' explanation of convection, and in part (b) the correct answer was 'conduction' or 'radiation'.

> In desert countries there is a shortage of fresh water.
> A scientist has suggested a new way of making fresh water from sea water.
> Electricity might be generated at the same time.
> The diagram shows the scientist's idea.



(a) Air rises inside the tower in a convection current.
Explain why convection happens.

_____

_____

_____ [2]

(b) Name **one other** way in which heat is transferred from place to place.

_____ [1]

The above question was manipulated by simply removing the diagram and preceding text. The intention was to test students' knowledge of convection in isolation of any context.

Removing the diagram did not make the question easier as we had thought; there was little difference between scores on the two different versions. The most obvious effect on part (a) was that more students left out the question when the diagram was missing (24.5% compared to 40 %). In interviews some students tended to perceive that having the diagram made the question easier, but performance did not change. Having the diagram seems to allow the students to guess an answer using the ideas in it and it may have acted as a reminder of the general idea of convection. 18.4% of students in version 1 tried to use the information in the diagram when answering the question when it was actually irrelevant. One student said at interview, "I tried to tie it in with the diagram and the whole tower thing but it was more or less a random guess". Other students were confused by the diagram and despite their commonly held stereotype that diagrams are relevant, recognised that it wasn't so in this case. The following statement was typical: "Although the diagram do make it clearer, I think this diagram is really a disappointment because it doesn't really say, it doesn't really explain why convection happens."

The marks scored on part (b) were also very similar for both versions. Students were not misled as much as we thought they would be by the diagram. Only 4.1% of those who had the diagram tried to answer using it.

The data that we gained in reference to this question were limited by the fact that in most of the schools this topic had not been studied recently.

Concluding comments

The examples we have shown illustrate the way in which students' expectations can influence responses in exam questions and hence students' marks. Students' experiences of classroom tests, textbooks and mock exams lead to the development of schemas that may later be activated when sitting an exam and can sometimes lead students to interpret questions or question requirements in a way not intended by examiners. In such cases the question will no longer be valid since the students' minds will not be "doing the things that we want them to show us they can do" and the examiners are no longer in control of what is being measured.

Assuming students' experiences of preparation for examinations can affect the way they answer a question, there is an element of exam technique being tested when the real intention of exams is to test conceptual knowledge, understanding and subject relevant skills. In some of the example questions in this study we managed to reduce the detrimental effect of expectations on validity to some degree. However, in only one question did the change help significantly. Our investigation of question 4 indicates that dramatic changes of topic are very unexpected and should probably be avoided whenever possible. In some cases students' expectations are too strong to suppress successfully during an exam, and our manipulation did not affect performance.

We could argue that whenever a question will contradict most students' expectations it should not be asked. However, there is a danger that this will reinforce stereotypes. This could result in inflexible skills that students can only use in stereotypical situations. However, with the threat to the validity that expectations may cause, perhaps the place for counteracting stereotypes is the classroom and not the exam. This leads to a further dilemma since it is hard to make something a necessary part of classroom activities if it is not being tested in an exam.

We have argued that expectations influence performance in exams on three different levels: subject, question and sentence level. Examiners can reduce the negative effects of expectations by using, for example, clear language, logical sequences of questions, relevant resources, and assessing one topic area at a time. It is important to avoid writing questions that contradict expectations and are vulnerable to misunderstanding. By avoiding misunderstandings as much as possible we can increase the validity of examination questions.

References

Ahmed, A. & Pollitt, A. (2000) *Observing Context in Action*. Paper presented at International Association for Educational Assessment Conference, Jerusalem, May 2000.

Ahmed, A. & Pollitt, A. (2001) *Improving the validity of contextualised questions*. Paper Presented at British Educational Research Association Conference, University of Leeds, September 2001.

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

Bramley, T., Hughes, S., Fisher-Hoch, H. & Pollit, A. (1998) Sources of Difficulty in Examination Questions: Science. *UCLES Internal Report*

Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J. & Cohen, G.M. (1997) Changes in Memory Awareness during Learning: The Acquisition of Knowledge by Psychology Undergraduates. *Journal of Experimental Psychology: General. 126*, (4), 393-413.

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

Frazier, L., & Rayner, K. (1982). Making and correcting errors in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*, 178-210.

Fisher-Hoch, H., Hughes, S. & Pollitt, A. (1998) Sources of Difficulty in Examination Questions: Geography. *UCLES Internal Report.*

Hughes, S., Fisher-Hoch, H. & Pollitt, A. (1998a) Sources of Difficulty in Examination Questions: Mathematics. *UCLES Internal Report.*

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.

Pollitt, A. & Ahmed, A. (1999) *A new model of the question answering process*. Paper presented International Association for Educational Assessment Conference, Bled, May.

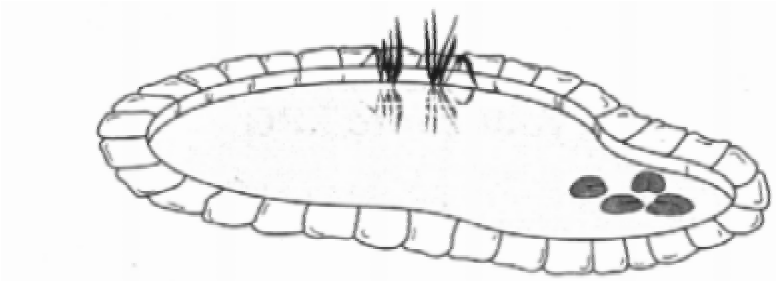
Pollitt, A. & Ahmed, A. (2000) *Comprehension Failures in Educational Assessment*. Paper presented at European Conference on Educational Research, University of Edinburgh, September 2000.

Rosch, E. (1978) Principles of categorization. In E. Rosch and B.B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Appendix 1 – Test Questions

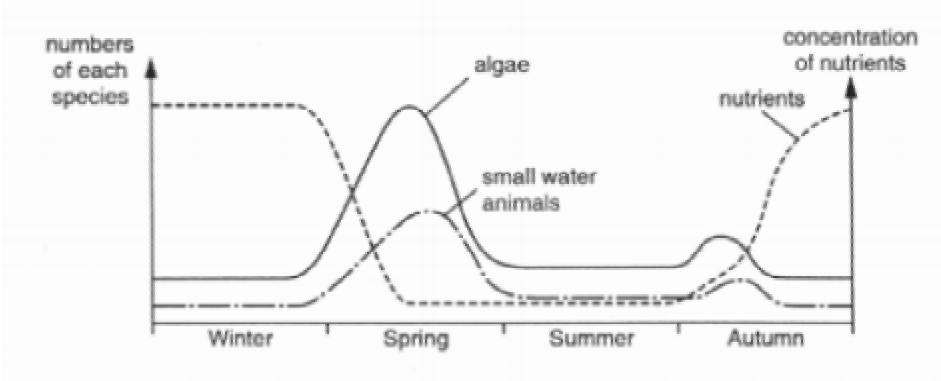
(Original version)

1. This question is about conditions in a garden pond.



The graph shows the levels of green algae, small water animals and soluble nutrients in a garden pond over a period of one year.

The small animals feed on the green algae.



(a) When is the concentration of nutrients in the pond highest?

……. [1]

(b) (i) Suggest two reasons for the rapid rise in the level of green algae in early spring.

1 …….

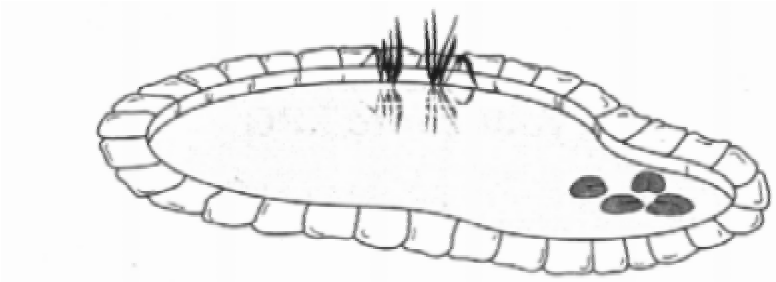
2 …….. [2]

(ii) Suggest why the level of green algae falls rapidly in late spring.

……. [1]

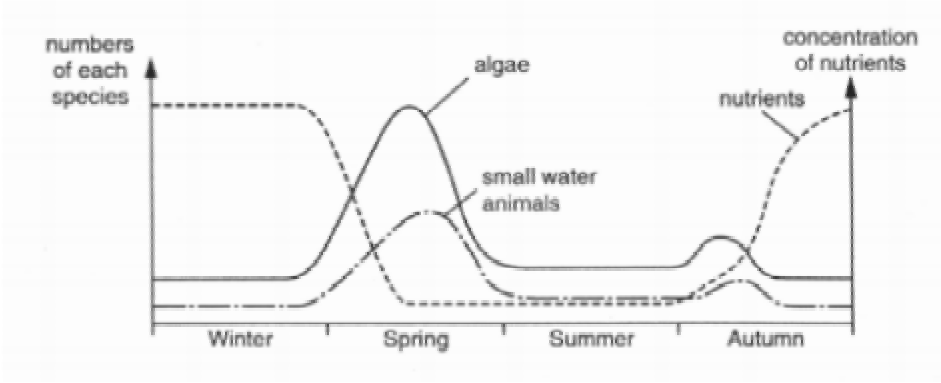
(Manipulated version)

1. This question is about conditions in a garden pond.



The graph shows the levels of green algae, small water animals and soluble nutrients in a garden pond over a period of one year.

The small animals feed on the green algae.



(a) **(i)** When is the concentration of nutrients in the pond highest?

……. [1]

(ii) Suggest why the level of green algae falls rapidly in late spring.

……. [1]

(b) Using your **own knowledge**, suggest two reasons for the rapid rise in the level of green algae in early spring.

1 …….

2 ……..

(Original version)

2. **(a)** Farmers add fertilisers to soil.

Write down one reason why.

…….. [1]

(b) Ammonium nitrate is a fertiliser.

It is made by reacting ammonia with nitric acid.

(i) Write down the word equation for this reaction.

……. [1]

(ii) A factory makes ammonium nitrate.

Write down **three** costs of making ammonium nitrate in the factory.

1…….

2…….

3……. [3]

(Manipulated version)

2. **(a)** Farmers add fertilisers to soil.

Write down one reason why.

…….. [1]

(b) Ammonium nitrate is a fertiliser.

It is made by reacting ammonia with nitric acid.

(i) Write down the word equation for this reaction.

……. [1]

(ii) A factory makes ammonium nitrate.

Write down **three** expenses for the factory of making ammonium nitrate.

1…….

2…….

3……. [3]

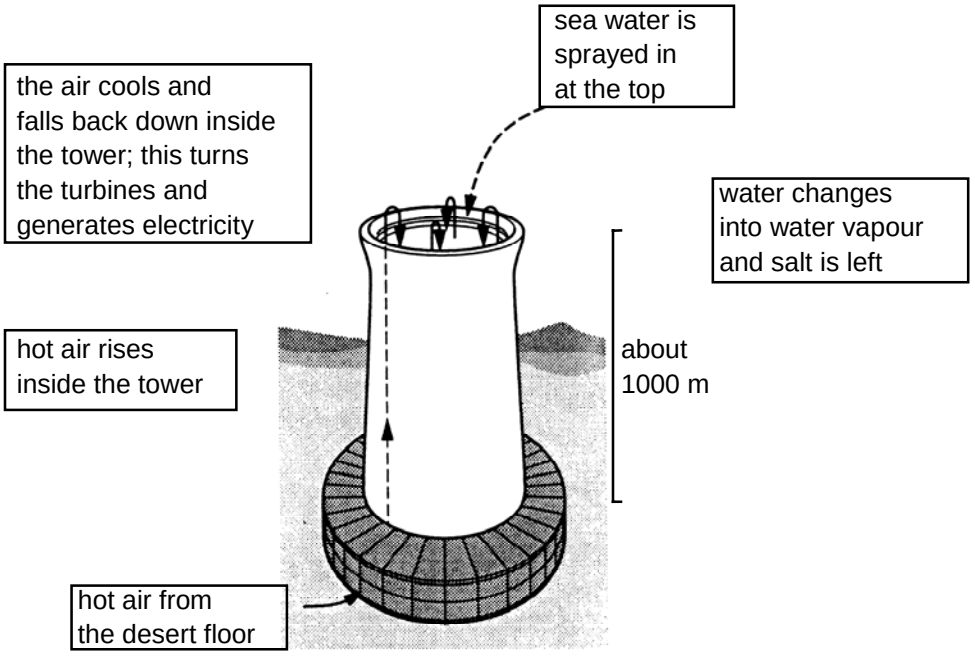
(Original version)

3. In desert countries there is a shortage of fresh water.

A scientist has suggested a new way of making fresh water from sea water.

Electricity might be generated at the same time.

The diagram shows the scientist's idea.



(a) Air rises inside the tower in a convection current.

Explain why convection happens.

……..

……

…….. [2]

(b) Name **one other** way in which heat is transferred from place to place.

…… [1]

(c) **(i)** Name the process which changes water into water vapour.

Choose from: **condensing evaporating freezing melting**

……[1]

(ii) What must you do to water to make this process happen?

……[1]

(Manipulated version)

3. (a) Heat can be transferred by convection.
Explain why convection happens.

……..

……

…….. [2]

(b) Name **one other** way in which heat is transferred from place to place.

…… [1]

(c) **(i)** Name the process which changes water into water vapour.
Choose from: **condensing evaporating freezing melting**

…… [1]

(ii) What must you do to water to make this process happen?

…… [1]

(Original version)

4. Joy plays a game.

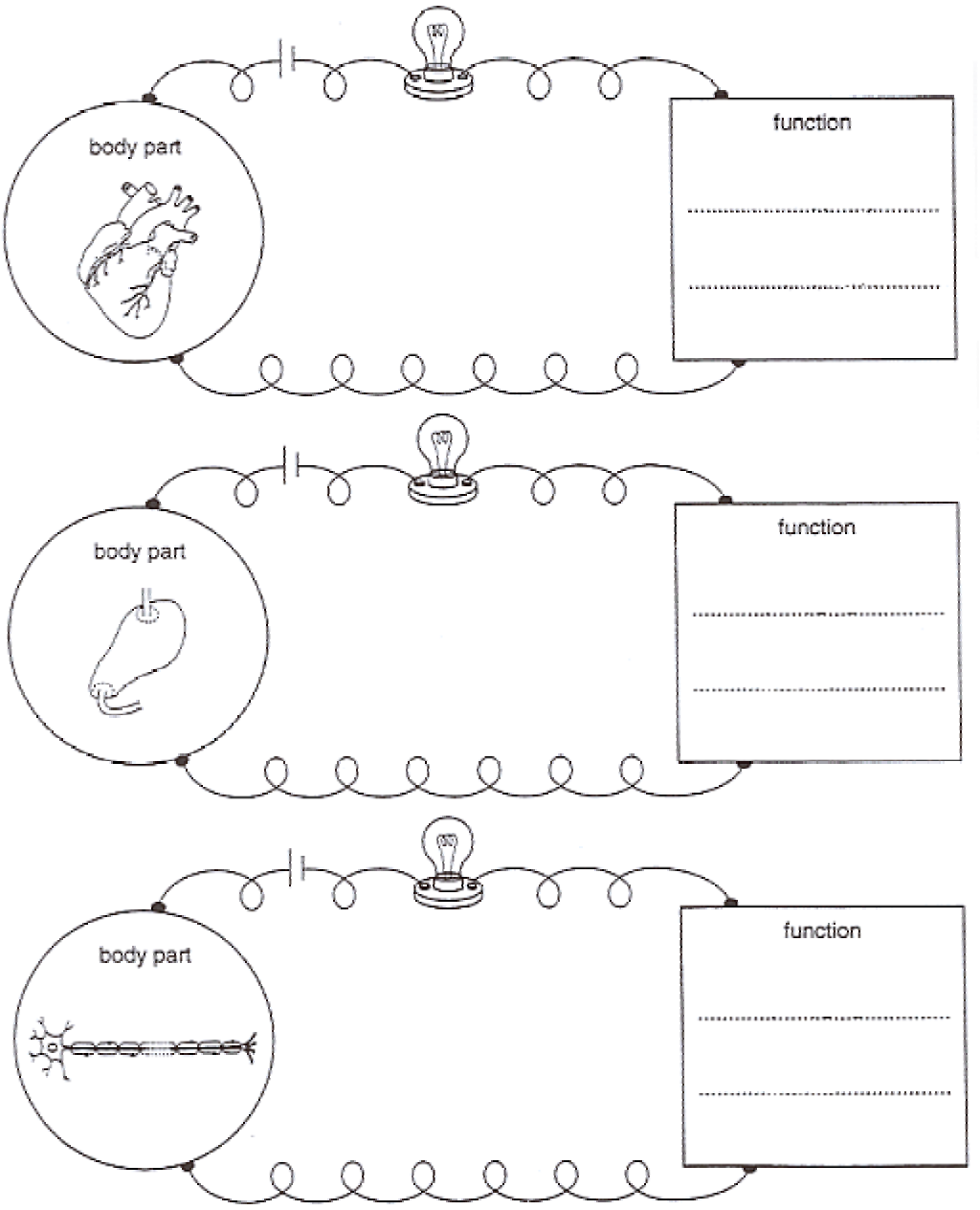
She has to match up parts of the body and their function.

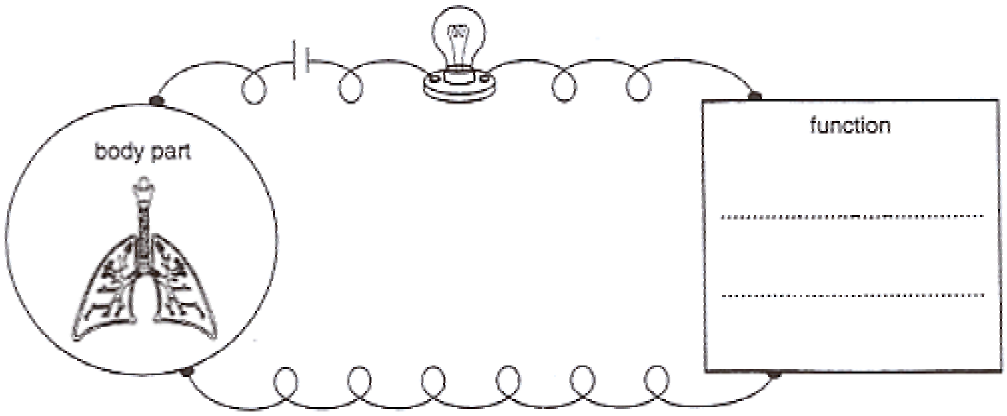
When she is correct a bulb lights up.

(a) Complete each box with the correct function.

Choose from:

carries nerve impulses **exchanges gases** **holds food** **pumps blood**





(b) Name one cell and one organ shown in these circles.

Cell ……..

Organ …….. [2]

(c) One day Joy finds that even with the right answer, the bulb does **not** light. Suggest two reasons why the bulb does **not** light.

1……

2…… [2]

(Manipulated version)

4. Joy plays a game.

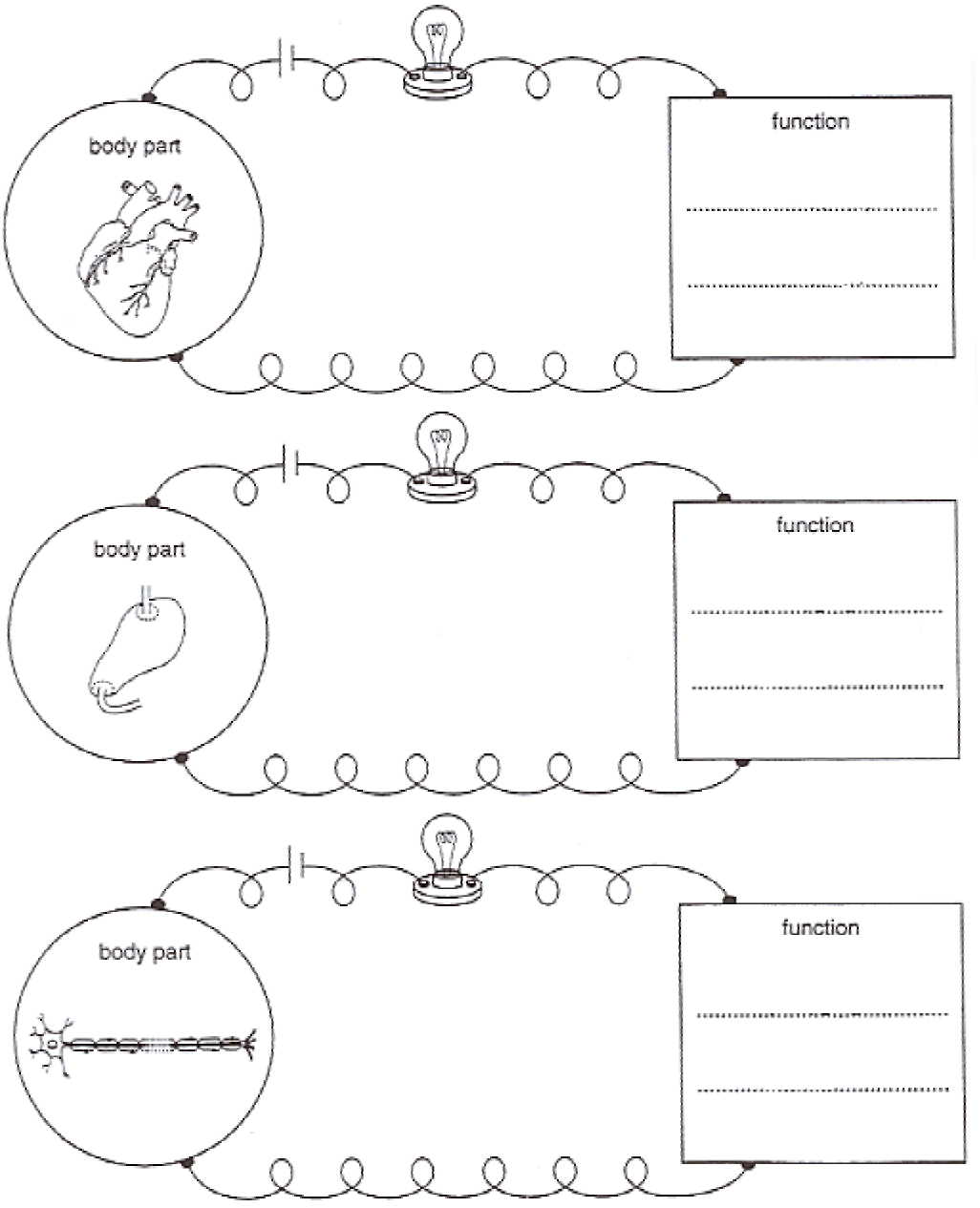
She has to match up parts of the body and their function.

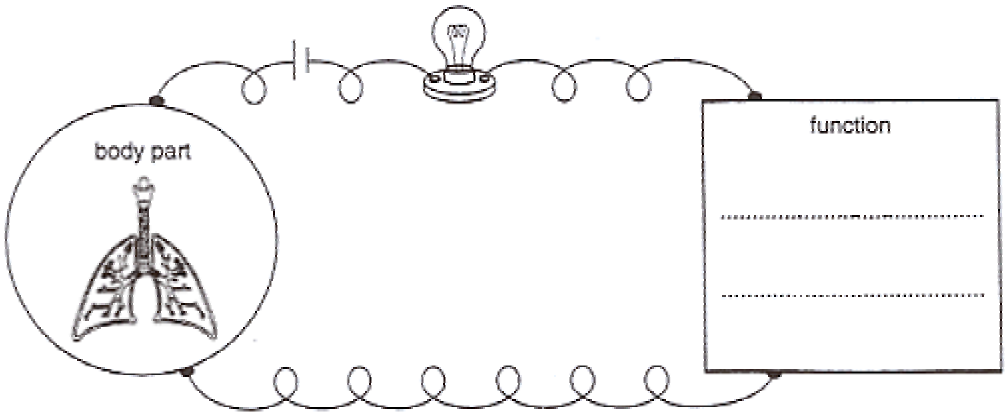
When she is correct a bulb lights up.

(a) Complete each box with the correct function.

Choose from:

carries nerve impulses **exchanges gases** **holds food** **pumps blood**



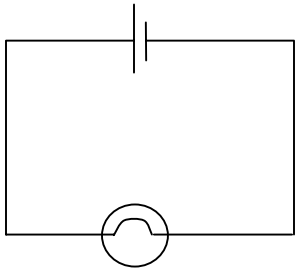


(b) Name one cell and one organ shown in these circles.

Cell ……..

Organ …….. [2]

(c) Look at the circuit diagram below.



The bulb usually lights but one day it does **not** light.

Suggest two reasons why the bulb does **not** light

1…….

2……. [2]

(Original version)

5. Read the following passage.

> Rachel is 15 and gets upset very easily. She is always arguing with her mother and older sister. She spends a long time in the bathroom or in her room trying out new hair styles and make-up. Rachel likes to go shopping with her friends and spends most of her pocket money on CDs. She also buys cigarettes and hides them at home. She has just bleached part of her hair. Her mother buys most of her clothes which are not fashionable enough for Rachel. When she goes out Rachel often borrows her older sister's clothes and make-up without asking and changes in the ladies' toilets. Rachel and her friends spend a lot of their time standing chatting to older boys and smoking in the shopping centre near the main music store.

(a) Use the information in the passage to suggest **one** cause of Rachel's arguments with her mother or sister.

……

……...[1]

(b) Suggest **one** way Rachel's mother could help to reduce the number of arguments.

……

……...[1]

(Manipulated version)

5. Read the following passage.

> Rachel is 15 and gets upset very easily. She is always arguing with her mother and older sister. She spends a long time in the bathroom or in her room trying out new hair styles and make-up. Rachel likes to go shopping with her friends and spends most of her pocket money on CDs. She also buys cigarettes and hides them at home. She has just bleached part of her hair. Her mother buys most of her clothes which are not fashionable enough for Rachel. When she goes out Rachel often borrows her older sister's clothes and make-up without asking and changes in the ladies' toilets. Rachel and her friends spend a lot of their time standing chatting to older boys and smoking in the shopping centre near the main music store.

(a) Use the information in the passage to state **one** cause of Rachel's arguments with her mother or sister.

……

……..[1]

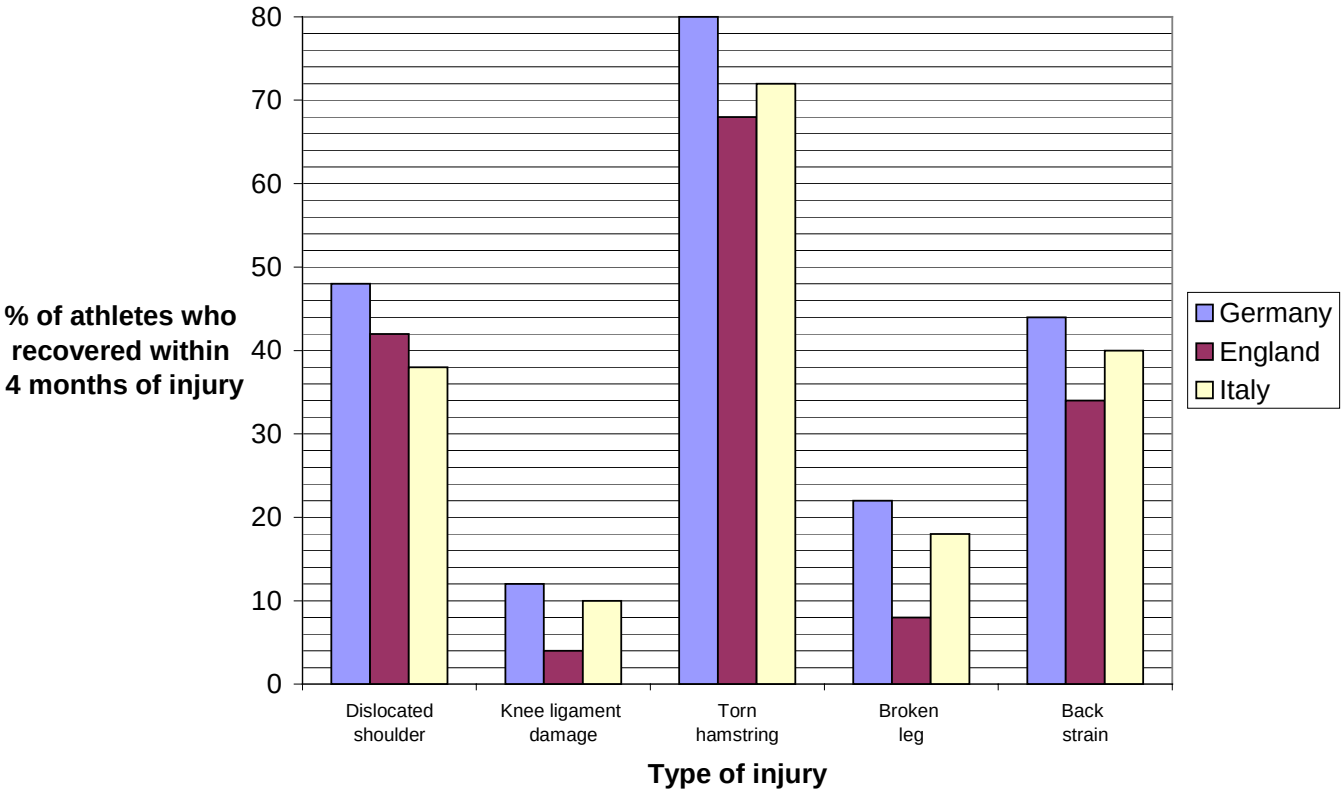
(b) Suggest **one** way Rachel's mother could help to reduce the number of arguments.

……

……..[1]

(Original version)

6. The bar chart shows the percentage of injured athletes in three countries who recovered within 4 months of injury.



(a) For which type of injury is the recovery rate the greatest?

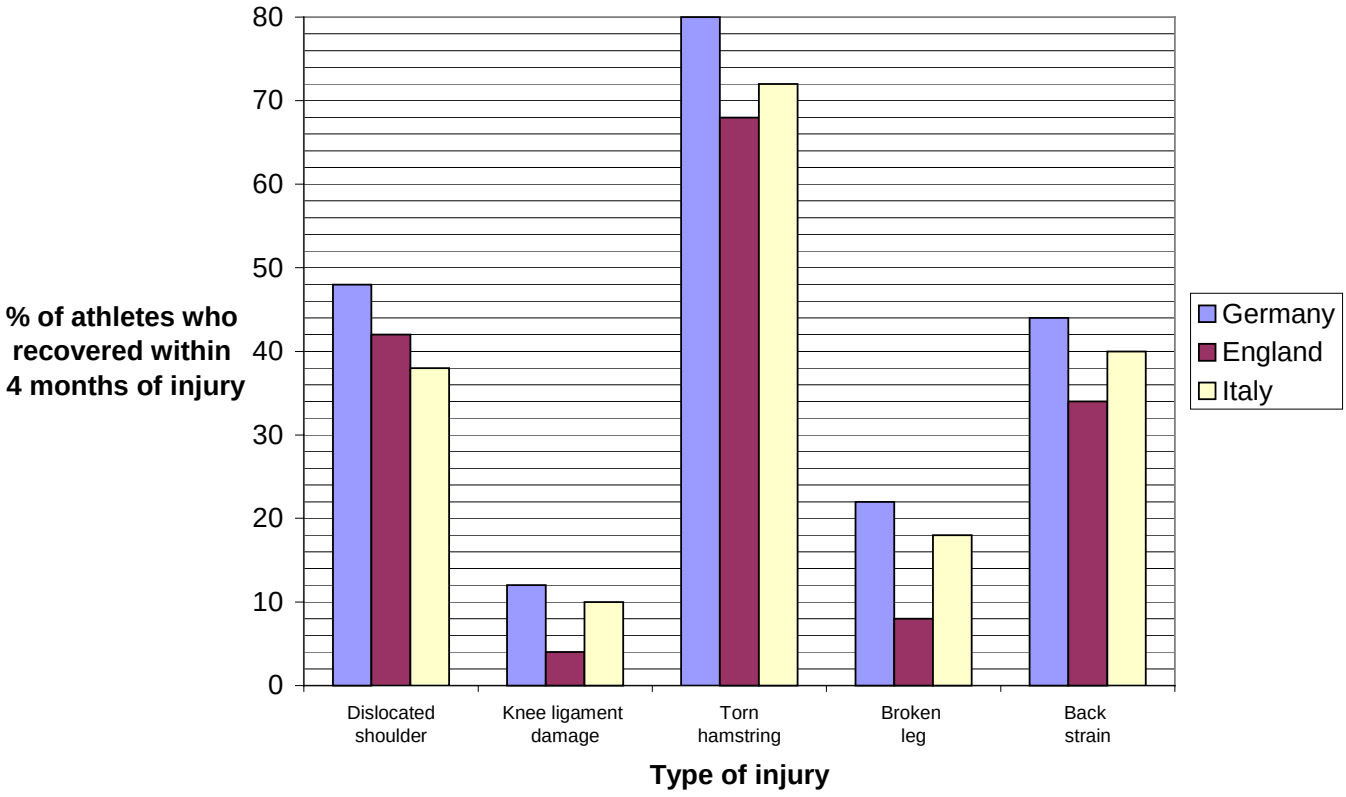
…… [1]

(b) What percentage of injured German athletes did not recover within 4 months of breaking a leg?

…… [1]

(Manipulated version)

6. The bar chart shows the percentage of injured athletes in three countries who recovered within 4 months of injury.



(a) For which type of injury is the recovery rate the greatest?

……. [1]

(b) What percentage of injured German athletes did **not** recover within 4 months of breaking a leg?

……. [1]

Appendix 1 continued – Mark Scheme

1. Pond

a. Winter (1) (ALLOW with late autumn)
Autumn + winter = 0

bi (Version 1) / b (Version 2)
increasing temperature / warmth (1)
increasing light / more daylight / more light (1)
nutrients = 0
IGNORE photo-synthesis

bii (Version 1) / aii (Version 2)
growth in number of (water) animals / eaten by (water animals) (1)

2. **Costs**

a. to provide nitrogen or nutrients/
to provide essential elements or minerals/
to increase crop yield /
to make plants grow faster/bigger/better/stronger Any 1
ALLOW to provide potassium / to provide phosphorus
ALLOW grows quicker/ helps plants grow
NOT healthier plants/ to provide nourishment or food/ to add nutrition

b (i) nitric acid + ammonia → ammonium nitrate (+ water) (1)
ALLOW = instead of arrow
NOT ammonia nitrate
ALLOW correct symbol equation (balancing not required)
ALLOW mix of formulae and words
NOT ammonium for ammonia

b (ii) Any **three**
(raw) materials / nitric acid / ammonia (1)
wages (1)
energy / electricity / gas (1)

plant costs (1)

ALLOW rates or taxes or loans / maintenance / waste removal / pollution control / research and development / safety / quality control

NOT just 'waste materials' / storage / transport / advertising / testing / pollution (on its own) / running the factory

3. **Tower**

a. (hot air) particles move faster / more spread out (1)
less dense / lighter (1)

b. conduction / radiation (1)

c (i) evaporating (1)

c (ii) heat it / (thermal) energy in / boil (1)

4. **circuits**

a. pumps blood (3) two or three correct = 2, 1 correct = 1
holds food
carries nerve impulses
exchanges gases

b. neuron / nerve cell(1)
heart / stomach / lung (1)

c. any two
bulb broken (1)
battery flat (1)
wire broken / circuit incomplete (1)
poor contacts (1)

5. angst

a. occupying bathroom
borrowing (aw) clothes without asking
borrowing (aw) makeup
dislike of clothes bought
bleaching hair Any 1
Reject: general reference to emotional changes
Reject: references to smoking unless in a 'finding out' context

b. consult with Rachel/go clothes shopping together/buy her own clothes
spend time talking/take an interest in clothes/makeup
compromise with Rachel about clothes/behaviour Any 1

6. graph

a. torn hamstring (1)
b. 78 (1)

Appendix 2 - Interview Questions

What did you think of the test?
Did you find any of the questions: particularly hard?
particularly easy?
Were there any questions where you didn't understand what was being asked?
Were there any topics that you have not studied/ couldn't remember?

QUESTION 1 - Garden Pond

(bi) How did you work out the answer to this question?
How did you know you had to use the graph to answer the question?

Manipulated version only:
What do you think is meant by 'use your own knowledge'?
Did you use the graph to answer the question?

QUESTION 2 - Costs

(bii) What do you think the question means?
(Tell them the mark scheme and see how they respond.)

FOR ALL QUESTION PARTS THAT WERE MANIPULATED:

Do you think your version was easier or more difficult? Why?

How did you work out the answer to this question?
What made you think you should answer in that way?

(If they made mistake that may have been a result of their expectations, tell them the right answer to try to prompt further comments)