Developing and testing a theoretical framework for assessing extended response questions in GCSE Science

Anne Mary Whitehouse

MA in Education (by research)

University of York

Education

February 2014

Abstract

This study aimed to develop, test, and validate a theoretical framework that could be used to write levels-based mark schemes for extended response questions in GCSE science. The work focused on questions which require students to give a scientific explanation or provide an argument. The development was informed by the work of researchers who have evaluated argumentation in the science classroom and also took into account the grade descriptors that are used to determine the cut scores for key grade boundaries during the awards process for GCSE science.

The framework was used to write mark schemes for five questions from the January 2012 GCSE Science examinations. The mark schemes were used to mark scripts (n = 19 to 26) from those examinations. The marks awarded were compared with those given by the examiners who originally marked the questions. To ensure the theoretically-based mark scheme could be used by others, three senior examiners also marked two of the questions.

Senior assessors (n = 12) for GCSE Science were asked, through an open response questionnaire, to comment on the framework and on its potential usefulness for writing levels-based mark schemes.

Comparison of marks awarded using the theoretically-based scheme with those using the original scheme showed that the two schemes produced similar rank orders (Kendall's coefficient τ = 0.61 to 0.83, n = 19 to 26). When other examiners used the theoretically-based scheme they awarded similar marks to those given by the researcher. These two outcomes suggest that the theoretically-based framework could be used for the proposed purpose. The senior assessors were generally positive about the usefulness of the scheme as a starting position for writing mark schemes and some recognised its potential to provide consistency of standards between different papers in the same session and across time.

Table of contents

List of tables

List of figures

Note on anonymity

This study used GCSE science examination papers, mark schemes and examination data. Since the specifications and examinations are cited in this report it is impossible to anonymise the awarding body.

To avoid the risk of individuals being identified by their roles, respondents to the questionnaire are all referred to as ‘assessors’.

Acknowledgement

This study would not have been possible without the source material and data provided by Cambridge Assessment, and by Oxford Cambridge and RSA Examinations.

Author's declaration

I certify that the work presented in this thesis was conducted by myself at the University of York, and I am the sole author.

This work has not previously been presented for any other award at any other institute. All sources are acknowledged as References.

Chapter 1 Introduction

1.1 The background for this study

Many people believe that it is important that school students should be able to express their ideas about science in connected prose (Wellington & Osborne, 2001). It is also argued that learning to put together an argument, based on evidence, will not only develop students' understanding of the science they are studying, but help them gain an understanding of the role of argumentation in making links between the conjecture of the scientist and the evidence that supports it (Newton, Driver, & Osborne, 1999). An understanding of the structure of a valid and convincing scientific argument and its role in the progress of science would be more likely to be taught in school if this understanding were to be explicitly assessed in public examinations. These ideas provided the stimulus for this study which began in 2011 at a time when new GCSE specifications for the sciences were introduced. For the first time the GCSE Science examiners were writing questions and mark schemes for which candidates would be required to write extended prose answers and also to show their ability to develop scientific arguments and explanations (Ofqual, 2009).

1.2 The place of argumentation in the National Curriculum

When the first National Curriculum for England and Wales was introduced in 1989, it included a requirement for students to be able to communicate their ideas about science in written prose (Department for Education and Science, 1989). By 1999 it was explicit in the National Curriculum documents that to achieve level 8 students must "communicate findings and arguments using appropriate scientific language and conventions, showing awareness of a range of views." (QCA, 1999, p. 76). In 2001 the Key Stage 3 National Strategy, a school support programme to raise standards in Key Stage 3, initiated a programme to improve students' literacy through a coherent programme across all subjects (Department for Education and Employment, 2001). In the following year, materials were published to support science teachers in developing writing skills in science including writing explanations, constructing arguments and drawing conclusions (Department for Education and Skills, 2002)

1.3 Assessment of argumentation in public examinations

In spite of the requirement of the National Curriculum for students to be taught to write explanations, construct arguments, and draw conclusions, there was, from 1989 to 2006, little explicit assessment of students' ability to carry out such tasks in external examinations. At various times between 1990 and 2006 a small proportion of the marks in GCSE examinations was given to the 'quality of written communication', typically 2 marks out of a total of 60 for the paper. These marks were usually allocated for the qualities of spelling, punctuation, and grammar. One mark would be allocated to each of two questions, for which the quality of the science content itself would typically score 2 or 3 marks. Between 2007 and 2011 the only place that the quality of written communication was explicitly assessed was in the context of teacher-assessed coursework, with no marks allocated for this in the examination papers.

New GCSEs in science were introduced in 2011 for first assessment in January 2012. The specifications and assessments were designed to match the GCSE Science Criteria (Ofqual, 2009). These stated that GCSE specifications in science must require learners to 'develop arguments and explanations, and draw conclusions using scientific ideas and evidence' (p. 6). This requirement implies that the assessments for the specifications must set questions that assess the candidates' ability to do these things. Ofqual also required GCSE Science examinations for the new 2012 specifications to include questions requiring extended written answers from candidates, with 6 marks being allocated to the marking of that answer in the mark scheme. Previously it was unusual to see individual parts of questions allocated more than 3 or 4 marks.

So examiners setting papers were required to do two new things – develop mark schemes for longer answers, and find ways of assessing candidates' abilities to evaluate information and develop arguments and explanations. These new question styles also posed a challenge to teachers in preparing students for the examinations. In the time leading up to the first assessment using this new style of questions, the professional development sessions provided by the awarding bodies included a focus on supporting teachers in this preparation.

The Twenty First Century Science course materials (Twenty First Century Science, 2006) and GCSE Science specifications (OCR, 2005) had made explicit the role of argumentation and evidence in bringing about acceptance of new science explanations. The examiners for this specification were already used to asking questions that tested candidates' understanding of the role of evidence. In 2011 they had the opportunity to set questions that required extended answers which could assess candidates' ability to develop an argument. It was through working, with these examiners in developing questions requiring levels-based mark schemes for the first time, and in developing support materials for teachers preparing students for this new question type, that led to the my interest in carrying out this study.

A study of the examiner reports and statistics from the first set of examinations showed that performance on these extended answers was variable (CA, 2012e; OCR, 2012a, 2012e). Whilst candidates did not avoid the questions, and indeed in many cases they filled the answer spaces, the standard of answers did not match the expectations of the examiners, as expressed by the mark schemes. Looking more closely at the mark schemes, showed that there was no obvious consistency in the requirements of the mark schemes across different papers; they did not appear to be based on any theoretical models, nor to be directly related to the grade descriptors given in the GCSE Science Criteria (Ofqual, 2009).

It is against this background that the development and evaluation of a theoretically-based marking framework for marking GCSE science questions that ask for arguments or explanations took place. A theoretical framework which could be used for such questions, across all examination papers might provide a way of maintaining more consistent standards between questions, papers and examination sessions. Sharing such a framework with teachers might help them to prepare students to writing better arguments and explanations.

1.4 An overview of this dissertation

Following this introduction, Chapter 2 outlines the process of writing, marking, and grading GCSE examinations in England. Much of the technical vocabulary used here is also included in a Glossary at the end of the dissertation.

Chapter 3 presents a review of some of the literature relating to the reliability of marking in public examinations, in particular considering how reliability of marking might be measured and whether reliability is affected by the type of question and mark scheme used. The second part of the chapter is a review of the literature on the subjects of teaching, evaluating, and assessing argumentation in school science. This chapter provides a context for the development and evaluation of the theoretically-based marking framework.

Chapter 4 identifies the research questions that the study aims to answer and outlines the stages of the work, including the rationale for the research strategy and methods used.

Chapter 5 then describes how a theoretically-based marking framework was developed and tested on a range of questions. In Chapter 6 the outcomes of using the theory-informed mark schemes are described. The framework was also presented to examiners and others involved in the examination process, and their responses are summarised and discussed in Chapter 7.

Finally, Chapter 8 reviews the conclusions of the study, considers what answers can be given to the research questions and discusses the implications for teachers, examiners, and science educators, with suggestions about further work that might be carried out following the ideas in this study.

Chapter 2 Context – the examination process

2.1 Overview

This chapter describes the process of setting, marking, and grading examinations in GCSE sciences. (Throughout this chapter where reference is made to GCSE sciences, this is intended to cover GCSEs in Science, Additional Science, Biology, Chemistry, and Physics.) The examination process is essentially the same for all subjects at GCSE and GCE level across all awarding bodies; it is overseen by the Office of Qualifications and Examinations Regulation (Ofqual) who regulate qualifications, examinations and assessments in England.

2.2 GCSE Criteria

The content of GCSE Science examinations is regulated by the GCSE subject criteria for science (Ofqual, 2009). These "set out the knowledge, understanding, skills and assessment objectives common to all GCSE specifications in science. They provide the framework within which an awarding body creates the detail of the specification" (p. 2).

The subject criteria include assessment objectives (AOs) which describe the things candidates will be required to do; awarding bodies must weight them within the ranges set out (Table 1)

Table 1 Assessment objectives GCSE Science (Ofqual, 2009, p. 7)

Assessment objectives		Weighting (%)
AO1	recall, select and communicate their knowledge and understanding of science	30-40
AO2	apply skills, knowledge and understanding of science in practical and other contexts	30-40
AO3	analyse and evaluate evidence, make reasoned judgements and draw conclusions based on evidence.	25-35

Each GCSE specification states the weightings that will be used to set the assessments, both for the examinations and for the internal assessments. Examiners setting papers must identify which AOs each question is testing (see 2.4.1)

The GCSE criteria also include grade descriptions, which "are provided to give a general indication of the standards of achievement likely to have been shown by candidates awarded particular grades." (p. 8). There are grade descriptors for each AO at grades A, C and F. These grade descriptors will be used when the grade boundaries for a particular paper are determined (see 2.6 Awarding grades).

2.3 Ofqual Code of Practice

The GCSE, GCE, Principal Learning and Project Code of Practice (Ofqual, 2011) describes the procedures that all awarding bodies must follow when producing, delivering, assessing, awarding, certificating and regulating qualifications, including GCSEs. The code is intended to ensure that assessments "be fit for purpose, command public confidence, be fair and accurate" (p. 4), and that standards will be maintained both over time and between and within awarding bodies.

The code describes the responsibilities of the personnel involved in the process of setting and awarding qualifications. Those of particular relevance to this study are identified here as they will be referred to later in this report.

The **chair of examiners** is responsible for maintaining standards across all the specifications within a subject area. For instance, if there is more than one suite of specifications for a subject, the chair must ensure that the standards set are the same for the two suites. The chair of examiners, or a deputy, chairs the question paper evaluation committee (QPEC), which meets to consider drafts of question papers and mark schemes to ensure that they are of high quality and match the specification.

The **chief examiner** for a specification is responsible for ensuring that all the units of assessment, both examinations and internal assessment, meet the requirements of the specification and that over a number of examination sessions standards are maintained and all aspects of the specification are assessed.

The **principal examiner** for a unit of assessment is responsible for setting the examination paper and mark scheme and standardising the marking of that paper.

Revisers provide written comments on early drafts of the paper and mark scheme and attend the QPEC.

The **scrutineer** checks the final draft of the paper and mark scheme, which have been produced by the Principal Examiner as a result of the QPEC discussions.

Assistant examiners are responsible for marking the candidates' work in accordance with the agreed mark scheme. Where there are large numbers of assistant examiners some examiners will also be **team leaders**, responsible for monitoring the marking of a group of examiners.

In this study the term 'assessors' will be used when referring to people carrying out all the roles described above and the term 'examiner' is used when referring specifically to people marking examination papers.

2.4 Producing question papers and mark schemes

The process of producing an examination paper is outlined Figure 1. This process is intended to produce examination papers of the highest possible quality. Papers must provide a valid assessment, that is, they must assess the things laid down in the specification; and they must also be reliable, so that the outcomes from any given assessment will be similar to those of papers set in previous years to similar cohorts of candidates.

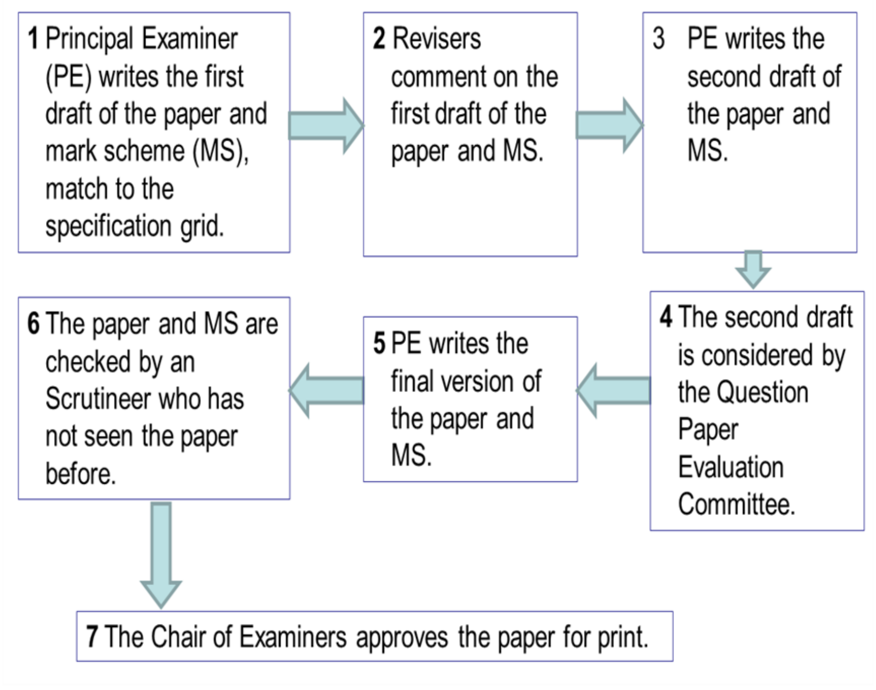

1 Principal Examiner (PE) writes the first draft of the paper and mark scheme (MS), match to the specification grid.
2 Revisers comment on the first draft of the paper and MS.
3 PE writes the second draft of the paper and MS.
4 The second draft is considered by the Question Paper Evaluation Committee.
5 PE writes the final version of the paper and MS.
6 The paper and MS are checked by an Scrutineer who has not seen the paper before.
7 The Chair of Examiners approves the paper for print.

Figure 1 The process of producing an examination paper

A high quality examination paper must also enable there to be discrimination between candidates. In the GCSE sciences there currently are two tiers of papers set for each unit of assessment. The higher tier paper is intended for those students expected to achieve a grade of C or above. The foundation tier paper is intended for those expected to get grades up to a C. So, for example, a higher tier paper must set questions that discriminate between candidates who can be described by the grade C descriptors and those who are performing at a lower level. But it must also identify those who fit the grade A description. Principal examiners are required to identify the area of the grade spectrum where they believe each question will provide discrimination (the target grade); this judgement is not currently based on any evidence of the difficulty of the question, but on the experience of the principal examiner (PE) and other members of the QPEC.

2.4.1 Specification grid

To help ensure the assessment matches the specification, the PE completes a grid similar to that shown in Figure 2. The specification grid shown is based on the one used to set GCSE Science papers for OCR. The paper has a maximum mark of 60 and the marks for each assessment objective must match the proportions laid out in the specification. There will be a row in the table for each part of each question. For each part-question the PE must provide a specification reference, identify the AO(s), and specify the target grade.

The target grades section shows the proportion of the marks that should be set at different levels of demand. Setting questions at different levels of demand should help the paper yield a spread of marks and so discriminate between candidates across the range of performance by the cohort taking the examination.

Question	Specification reference	Assessment objectives			Total mark	Target grades			Objective marks	Quantitative skills
		AO1	AO2	AO3		Standard demand	High demand			
						D, C	B, A	A*		
		24±2	27±2	9±1	60	32	20	8	Max 24	12-15

Figure 2 Example of a specification grid for a higher tier paper

To ensure that there are enough questions that require prose answers, there is a maximum number of marks allowed for ‘objective’ questions. These are considered by Ofqual to be questions for which only 1 mark is awarded; normally the answers to these questions are unambiguous, and the mark scheme can easily list acceptable answers. There is also a requirement to ensure that science papers include questions that require candidates to demonstrate quantitative skills; the grid in Figure 2 shows a target of 20-25% of the total marks.

2.4.2 Mark schemes

Mark schemes are written at the same time as the questions – the PE must show the expected answers and how marks will be distributed where there is more than one mark for a question. As well as objective mark schemes, described above, there are two other categories of mark scheme.

Points-based mark schemes are used for structured questions where the answers may be given in short sentences, up to one or two paragraphs, or a diagram or graph. Marking questions of this type usually involves identifying the relevant points in the candidate's work, and matching them to a list of acceptable points in the mark scheme. There may be a one-to-one correspondence between marks for the question and points listed in the mark scheme, or the mark scheme may give more alternatives than there are marks. The examiner identifies acceptable points up to the maximum number of marks available.

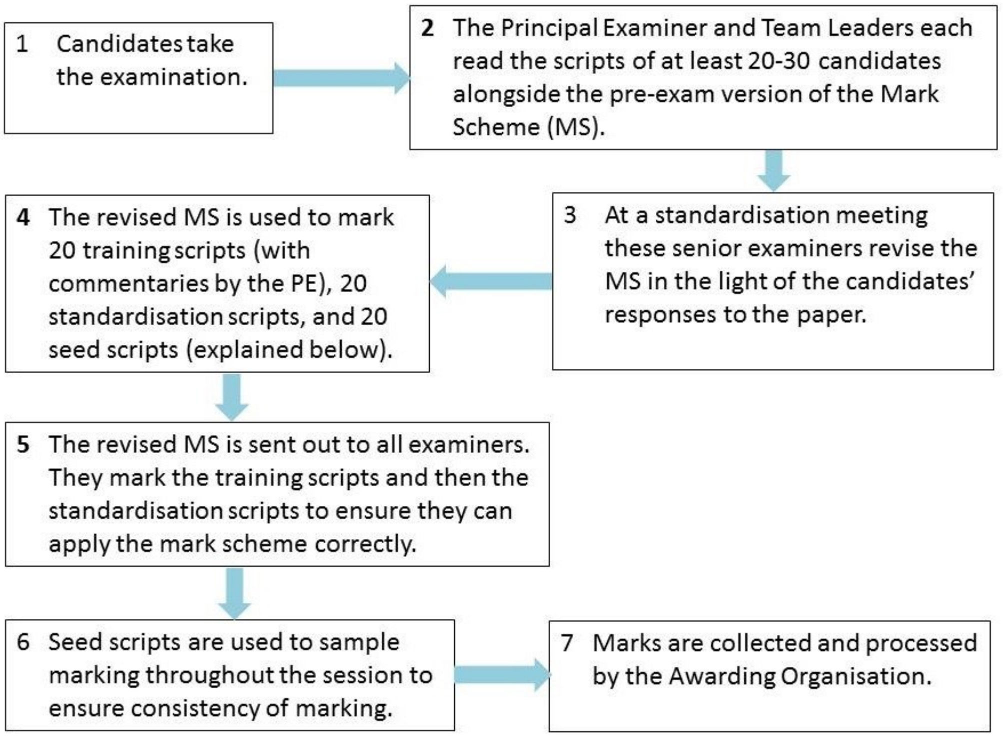
Levels-based mark schemes are used to mark longer prose answers – from one or two paragraphs up to extended essays. The mark scheme describes a number of levels of response, each with an associated band of marks. The description for each band will identify the criteria a candidate's response needs to match to be in that band. Normally examiners apply a principle of 'best fit' to decide the mark to award. This type of mark scheme is also called a level of response (LOR) mark scheme, or a banded mark scheme.

For specifications based on the 2009 GCSE subject criteria for science, Ofqual required the examinations to include questions that asked candidates to write prose answers worth 6 marks. These answers were to be marked using levels-based mark schemes. In these mark schemes, three levels of response are described, each associated with a band of marks (1–2, 3–4, 5–6). In the examination papers used in this study each 60 mark paper has three questions with levels-based mark schemes.

The GCSE science examinations based on the 2009 criteria, and taken for the first time in January 2012, were the first to include questions with levels-based mark schemes; previously papers had not included part-questions worth more than 4 marks and all questions other than objective questions were marked using a points-based mark scheme.

2.5 Marking GCSE examination papers

Figure 3 outlines the processes that take place from when the candidates take the examination, to having a list of marks for all candidates. This process is designed to ensure that marking is as reliable as possible – so that no matter which examiner marked a script, the same mark would be awarded. For this to be the case all examiners must have a common understanding of the mark scheme and be able to apply it in the same way.



1 Candidates take the examination.

2 The Principal Examiner and Team Leaders each read the scripts of at least 20-30 candidates alongside the pre-exam version of the Mark Scheme (MS).

3 At a standardisation meeting these senior examiners revise the MS in the light of the candidates' responses to the paper.

4 The revised MS is used to mark 20 training scripts (with commentaries by the PE), 20 standardisation scripts, and 20 seed scripts (explained below).

5 The revised MS is sent out to all examiners. They mark the training scripts and then the standardisation scripts to ensure they can apply the mark scheme correctly.

6 Seed scripts are used to sample marking throughout the session to ensure consistency of marking.

7 Marks are collected and processed by the Awarding Organisation.

Figure 3 The marking process for on-screen marking

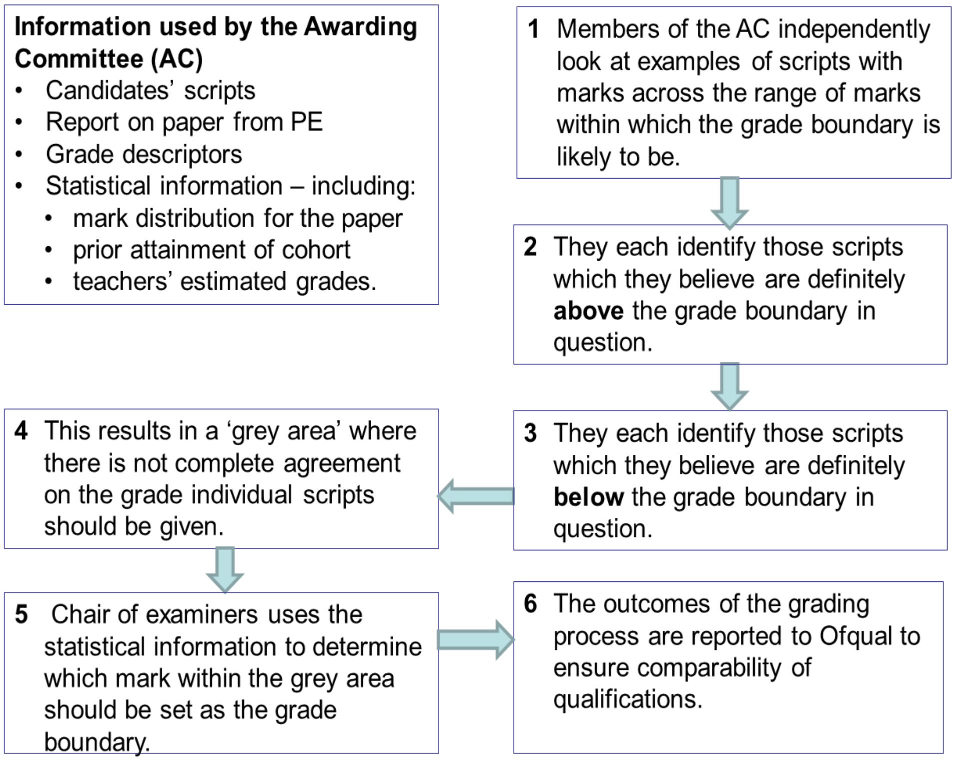
Senior examiners (principal examiners and team leaders) apply the mark scheme that was approved at the end of the question paper development process to a sample of candidates' scripts. This stage (box 2 in Figure 3) is to try to ensure that the mark scheme covers the range of responses that candidates have given. These senior examiners meet for a standardisation meeting (box 3), where the mark scheme is finalised and additional guidance is added to aid examiners in making decisions when awarding marks. These senior examiners also agree the marks on the scripts that will be used for training, for standardisation, and for sampling the marking of examiners. The principal examiner writes commentaries to go with the training scripts to ensure that all examiners come to a good understanding of the mark scheme.

Assistant examiners are required to mark the 20 training scripts and they then mark 10 'standardisation scripts' that will be used to check that they are marking to the required standard. Team leaders check that their team are giving the same marks for the standardisation scripts as was agreed by the senior examiners. If the examiner is marking accurately he/she is allowed to proceed to marking live scripts; if not the team leader (TL) will give feedback and then ask for a further 10 standardisation scripts to be marked; if these are satisfactory the examiner can proceed to marking. The quality of all examiners' marking is monitored throughout the marking process. Much of the marking is now done 'on screen', 60-88% of the papers from the main awarding bodies were marked on-line in 2012 (Ofqual, 2013c). During on-line marking each batch of 20 scripts that an examiner marks will include a 'seed script'. Any discrepancies between the examiner's mark for answers in the seed script and the agreed marks will be reported automatically to the TL. This enables the TL to easily monitor an examiner's marking throughout the marking process; examiners who are not marking to the correct standard may be withdrawn from the process and the scripts they have marked will be remarked by another examiner. In subjects where examiners mark the paper scripts they send samples of their marking to the team leader who checks they are continuing to mark to the agreed mark scheme.

2.6 Awarding grades

The Code of Practice (Ofqual, 2011) describes in detail how awarding bodies should determine the grade boundaries for each assessment. This process is outlined in Figure 4. Whilst the awarding bodies are responsible for determining the grade boundaries they do this within constraints laid down by Ofqual which are intended to ensure that " the qualifications are comparable – a student should get the same grade for their work whichever exam board they use – and the grade standards set are appropriate" (Ofqual, 2013a).

Steps 1-3 in Figure 4 rely on the judgement of the members of the awarding committee and in particular those examiners who are familiar with the examination paper, mark scheme and the performance of candidates on that paper. This aspect of the awarding process is referred to as 'construct-referenced' assessment (P. Black, 1998), decisions made solely on the basis of the judgement of examiners is open to criticism, particularly if the proportion of candidates achieving a particular grade changes from one session to the next. In practice the decisions about where the grade boundary is set is also informed by the statistical information for the paper (step 5) and also similar data for the same paper at previous sessions.



Information used by the Awarding Committee (AC)

- Candidates' scripts
- Report on paper from PE
- Grade descriptors
- Statistical information – including:
  - mark distribution for the paper
  - prior attainment of cohort
  - teachers' estimated grades.

1 Members of the AC independently look at examples of scripts with marks across the range of marks within which the grade boundary is likely to be.

2 They each identify those scripts which they believe are definitely **above** the grade boundary in question.

3 They each identify those scripts which they believe are definitely **below** the grade boundary in question.

4 This results in a 'grey area' where there is not complete agreement on the grade individual scripts should be given.

5 Chair of examiners uses the statistical information to determine which mark within the grey area should be set as the grade boundary.

6 The outcomes of the grading process are reported to Ofqual to ensure comparability of qualifications.

Figure 4 The process of awarding grades (Ofqual, 2013a)

When a new specification is examined for the first time, there may be less statistical information available; the cohort may have changed with the change in specification, and the style and demand of the question paper may have changed. To support the grade boundaries set on the new science qualifications in summer 2012, chairs of examiners were asked to provide samples of scripts at each grade boundary and give a narrative account of how the work matched the grade descriptions (Ofqual, 2012). In summer 2013 chairs of examiners had to confirm to Ofqual, "whether or not the work seen in scripts sufficiently matched the grade descriptions." (Ofqual, 2013d).

2.7 Examination statistics

Statistics for the paper as a whole have always been used as part of the awarding process. Comparing data about the marks awarded with those awarded in previous years gives an indicator of whether the demand of the paper was similar or not (assuming the cohort is similar).

On-screen marking of exam papers has enabled awarding bodies to collect far more data about the examinations and candidates than was previously possible. Data can be collected for each part of each question. This can reveal whether particular questions were missed by many candidates – perhaps because they were too demanding or because of where they appeared on the page – or maybe if the paper was too long, the last question may not have been completed. Before the collection of this item level data (ILD), a principal examiner may have reported that there seemed to be questions that were omitted or performed badly, but that would have been based on the impressions of markers, rather than on any systematic collection of data.

It is possible for senior examiners to see how each individual question performed in the examination; ILD provides information about the facility of the question (the mean mark for the question as a proportion of the maximum mark) and whether it discriminates between good candidates and weak candidates. Discrimination is measured by calculating the correlation between candidates' mark for the question and their overall mark for the paper, in other words, do candidates who do well on the paper overall do well on this question? A weak or negative correlation would raise concerns about how candidates were answering the question – for example there may have been an ambiguity which confused more able candidates but was missed by weaker candidates. Item Characteristic Curves (ICC) (Figure 5 and Figure 6) show how the facility of a question varies for different parts of the cohort (Elliott & Johnson, 2007). This is an effective way of displaying both the demand and the discrimination of a question. A steep downward slope indicates good discrimination, whereas a shallow slope indicates that there is little difference in performance between different parts of the cohort.

In the example in Figure 5 the facility for A grade candidates is about 0.9 – in other words most of those candidates scored most of the marks. At grade E the facility was about 0.2 – on average the E candidates only scored one-fifth of the marks on this question. This question discriminated well between candidates, and there was little difference between the performance of boys and girls.

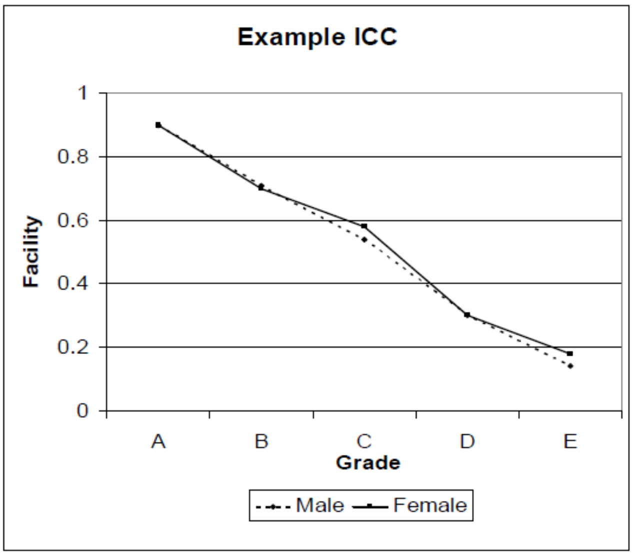


Figure 5 Sample Item Characteristic Curve (Elliott & Johnson, 2007. p5)

The sample curve in Figure 6 tells a different story; whilst there is discrimination across the grades, even the best candidates only scored an average of half marks on the question.

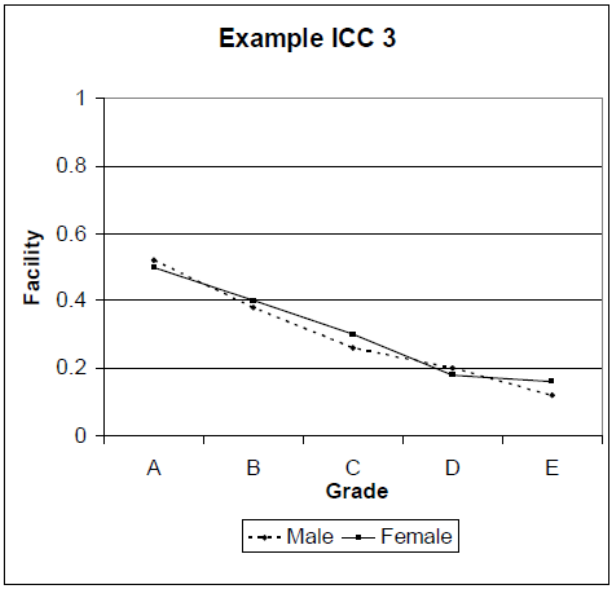


Figure 6 Sample Item Characteristic Curve (Elliott & Johnson, 2007. p5)

The awarding bodies also make data about the papers available to centres (schools and colleges). Teachers in centres can obtain detailed information data about how their candidates performed across different components of an assessment – at the level of individual candidates (OCR, 2013a).

2.8 Reports to centres

At the end of each examination series the awarding bodies publish reports to centres. The purpose of the reports is summarised by the statement which is included at the beginning of all the OCR reports:

> This report on the examination provides information on the performance of candidates which it is hoped will be useful to teachers in their preparation of candidates for future examinations. It is intended to be constructive and informative and to promote better understanding of the specification content, of the operation of the scheme of assessment and of the application of assessment criteria. (2012s, p. 2)

Each principal examiner writes a question-by-question report on the performance of candidates. Ideally they include information about the strengths and weakness of the cohort for each part, in particular identifying common misunderstandings revealed in answers to questions. Many teachers use these reports to inform their teaching; this exemplifies one of the ways in which assessment can have a backwash effect on what is taught in the classroom (Biggs, 1998).

Chapter 3 Literature review

3.1 Introduction

This study aims to develop a theoretical framework for the marking of GCSE examination questions that ask for an argument or explanation to be presented. Any mark scheme developed must be at least as good as the mark schemes used currently. The first section of this chapter reviews the evidence related to factors that affect the reliability of mark schemes and methods of quantifying the reliability. Section 3.3 outlines some of the reasons why teaching argumentation is considered to be an important element of science education; section 3.4 describes the way some researchers have evaluated argumentation in science lessons. Finally there is a brief consideration of the assessment of argumentation in GCSE science.

3.2 The reliability of marking of public examinations

GCSE examinations are a high stakes assessment – the outcomes matter to students who need particular grades to progress to the next stage of the education; they matter to schools and colleges who are judged by the progress of their students, shown by their position in league tables; and the outcomes are also important to teachers, who may be judged by the performance of their students (Baird, Ahmed, Hopfenbeck, Brown, & Elliott, 2013). These are good reasons to make every effort to ensure that the marking of the assessments is reliable, that is, “candidates should receive marks as close to their correct, ‘true’ scores as is possible, and that this should be the case no matter who marks their work.” (Ofqual, 2013c, p. 3). Although in an ideal world there would be no discrepancies between the marks that would be awarded by different examiners for a particular answer, when questions are not objective there is an element of judgement involved and sometimes there are likely to be differences in the mark awarded. The awarding bodies, through the Joint Council for Qualifications, have agreed that the tolerance on written papers should be ±6% of the paper total (rounded up to the next whole mark) (OCR, personal communication, 22nd July, 2013).

The concerns about reliability of examinations are demonstrated by the fact that there have been three official reviews of reliability, commissioned by the National

Assessment Agency and by Ofqual, in the space of eight years (Meadows & Billington, 2005; Ofqual, 2008, 2013c). The research programme of Ofqual also resulted in a special issue of Research in Education (Baird & Black, 2013), which covers a much wider range of issues related to reliability of public examinations than those considered in this study.

Of course the awarding bodies have a keen interest in making marking as reliable as possible and much of the research referred to in this chapter was carried out by staff working for the research groups of either Cambridge Assessment or AQA (Baird, Greatorex, & Bell, 2004; B. Black, Sütő, & Bramley, 2011; Bramley, 2008; Chamberlain, 2008; Meadows & Billington, 2005; Sütő & Nádas, 2008; Sütő, Nádas, & Bell, 2009).

3.2.1 Factors affecting reliability of marking - an overview

There is an inevitable tension when researching factors affecting marking reliability between using evidence from the marking of live papers, where it is difficult to control variables but it is an authentic environment, or alternatively generating evidence in an experimental situation where some variables can be controlled much more easily, but examiners may not experience the same pressures. An experimental study of marker reliability by Chamberlain (2008) included a questionnaire sent to all examiners, designed to compare their conscientiousness when marking for the experiment with that when marking live papers. It was found that a small proportion of the examiners acknowledged that they did not respond in the same way to the task as they would have when marking live papers.

There has been much research into factors affecting the quality of marking and the terminology used to describe the variability in marks awarded for a particular piece of work varies. Newton (2009), writing about reliability of national curriculum testing at key stages 2 and 3, made a distinction between reliability and accuracy of assessment. He stated that reliability "refers to the consistency of outcomes that would be observed from an assessment process were it to be repeated" (p. 183). The reliability of a student's grade is affected by factors such as the year in which they took the test, the version of the test, or the marker. Newton pointed out that there are also non-random factors that could result in students not receiving the correct mark that may be to do with different groups of students responding to a question in different ways, perhaps due to the language used or the contexts for questions; these factors may introduce systematic errors, which would make a difference to some parts of the cohort. It is these types of errors that he described as inaccuracies. These systematic errors are not due to the marking procedures, and in principle should be identified when the question papers are written.

Other researchers have used the term accuracy when they are looking at factors that affect whether the marker is awarding the same score as the principal examiner (or equivalent) would have done for the same question (Bramley, 2007; Johnson, Penny, Gordon, Shumate, & Fisher, 2005; Sütő & Nádas, 2008; Sütő et al., 2009). In other studies of these ideas, authors have used the term marker reliability (Baird et al., 2004; Massey & Raikes, 2006), marker agreement (Bramley, 2007), or marking quality (Bell, Bramley, Claessen, & Raikes, 2007). Although the researchers used different terms, all these studies were looking at the factors that might affect the mark or grade awarded for a particular response; this is the sense in which marker reliability is used in this study. This review will focus on issues of marker reliability, rather than the concerns Newton described as inaccuracies (2009).

Research shows that the reliability of marking depends on a whole range of factors. Sütő and Nádas suggested that factors can be divided into two groups – the demands of the marking task and the markers' expertise (2008). Figure 7 shows some of the factors that they suggested contribute to marking reliability. Whilst the diagram includes many possible factors, it does not take into account that the way the marking process is carried out may itself be a factor. Spear (1997), carried out a study of the effect on the mark awarded of the order of the scripts marked by teachers (not necessarily examiners) she suggested that the order in which questions are marked can have an effect on the mark awarded – a good answer that follows a poor answer may be credited more highly than if it had preceded the poor answer. There may be differences in marking reliability depending on whether it is carried out on paper or on-screen; marking on-screen also makes it easy to divide up an examination paper so that different examiners mark different parts of the paper. These issues are of growing relevance as more marking is now carried out on-screen (Ofqual, 2013c; Tisi, Whitehouse, Maughan, & Burdett, 2013).

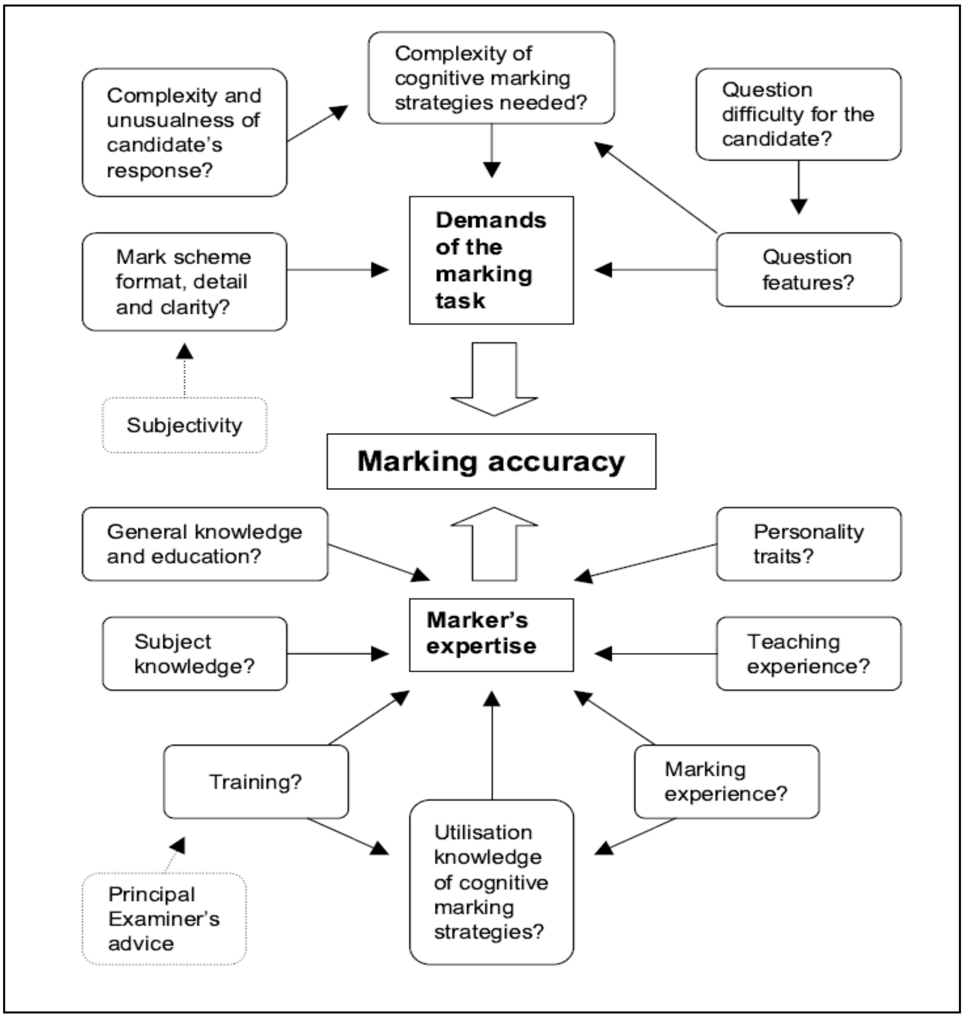

Complexity and unusualness of candidate's response?
Complexity of cognitive marking strategies needed?
Question difficulty for the candidate?
Mark scheme format, detail and clarity?
Demands of the marking task
Question features?
Subjectivity
Marking accuracy
General knowledge and education?
Personality traits?
Subject knowledge?
Marker's expertise
Teaching experience?
Training?
Marking experience?
Utilisation knowledge of cognitive marking strategies?
Principal Examiner's advice

Figure 7 Diagram summarising some key factors that potentially contribute to marking accuracy, indicating the main relationships hypothesised among them (Sütő & Nádas, 2008, p. 481)

3.2.2 Quantifying reliability of marking

Researchers have used a variety of statistical measures to quantify the level of agreement between markers. Bramley (2007) suggested a simple measure of agreement on marking of objective questions (P_0) which is the proportion of questions where the markers award the same mark as the principal examiner (PE). He suggested that P_0 can be said to describe the accuracy of the marking as the mark awarded by the PE on an objective question should be the 'true' mark. However this measure does not give any indication of the size of any discrepancies and whether a marker tends to be awards higher or lower marks than the PE. (For objective questions where no judgement is required, any differences between the mark awarded by the marker and the PE must be due to marker errors.) For a particular question, the standard deviation of the discrepancies would indicate the spread of differences for that question, and the sign of the mean of the differences would show whether the marker was generous or harsh.

For questions which are not objective, where there is some degree of examiner judgement involved, such as levels-based mark schemes (see 2.4.2), Bramley suggested that there is not one ‘true’ mark for the question, and the term ‘agreement’ would be better than ‘accuracy’. In this situation agreement (P_{agr1}) is the proportion of cases where the discrepancy between marks awarded by the marker and the PE is at most ±1. The significance of this discrepancy will depend on the maximum mark for the question. Again the standard deviation and mean of the differences would give further information. Bramley suggested that a scatter plot of differences may be useful to indicate any variability in the harshness of the marker – for instance if they were generous with good candidates and harsh with weak candidates, the mean difference may be small, and not reflect the variability. He pointed out that a simple calculation of correlation would show covariation but not agreement; there may be a good correlation even though agreement P_{agr1} is low.

Massey and Raikes (2006) studied the variability in marks awarded by three or four examiners for the same questions on 300 scripts for each of five subjects. They used the intraclass correlation coefficient (ICC), which reflects the degree of agreement amongst the examiners, rather than comparing assistant examiners with a ‘gold standard’ principal examiner.

Sütő and Nádas (2008) used similar measures of accuracy to those proposed by Bramley (2007), that is accuracy P_0, mean actual difference, and also the mean absolute difference.

3.2.3 The effect of examiner training and standardisation on reliability of marking

Traditionally awarding bodies have recruited appropriate subject teachers to mark examination papers; for many teachers this is a way of earning an additional income and at the same time learning more about the assessment of the subjects they are teaching in school (OCR, 2013b). According to the Joint Council for Qualifications (2013) most examiners are still teachers, whilst two of the awarding bodies, AQA

(2013) and Edexcel (2013a) stipulate that to be an examiner applicants must have teaching experience, OCR (2013b) do not. Whether examiners have teaching experience or not, there will always be some training before they mark 'live' examination scripts.

When training examiners to use levels-based mark schemes it is common practice to provide examples of candidates' answers that match the level descriptors (see section 2.4.2). Baird, Greatorex and Bell (2004) carried out a study to find the effectiveness of using exemplar scripts to improve the reliability of marking. They divided a team of 45 examiners into three groups to mark a GCSE English literature examination. The essays in the paper were marked using a levels-based mark scheme with six bands. One group of examiners was provided with two exemplar scripts at the bottom of each mark band; another group was given four exemplar scripts that were intended to be prototypical of each of three of the mark bands, being set in the middle of the three bands; the third, control, group did not receive any exemplar scripts. The exemplar scripts were used to train the examiners in the application of the mark scheme. The two groups received the exemplar scripts unmarked; they marked them using the mark scheme and then returned them to the principal examiner for comment. The principal examiner gave them feedback as well as telling them the 'correct' mark (i.e. the mark the principal examiner awarded). All 45 examiners then marked the same 150 scripts.

The researchers compared the marks awarded by the three groups with those awarded by the principal examiner. Perhaps surprisingly the most accurate marking was by the group who received no exemplar scripts. The group who received prototypical (mid-band) exemplar scripts were the harshest markers, perhaps thinking of the exemplar scripts as ones which had only just reached the standard awarded, and hence setting a higher standard than was intended. The group of examiners who received exemplar scripts which were just within the marking bands were slightly generous, but not significantly so.

Does this mean that exemplar scripts are not helpful? This study was carried out with experienced examiners who had marked the other English literature paper set at the same session as this experimental paper, so these examiners would already have in mind the standards expected for each band. This suggests that these experienced examiners already belong to the sort of community of practice described by Wenger (2000) as being ‘bound together by their collectively developed understanding’ (p. 229). For new examiners the exemplar scripts might prove an important ingredient in gaining that understanding.

Baird et al (2004) also carried out a study to find out the effectiveness of coordination meetings in increasing the accuracy of marking. In this study 36 experienced examiners were asked to mark GCSE History papers; they had all marked paper 1 at the summer examination session and in this study they were to mark paper 2. The examiners were placed into one of three matched groups. The control group were asked to mark scripts using just the mark scheme and exemplar scripts which had been marked by the principal examiner. The second group attended a coordination meeting which had a hierarchical format – the principal examiner explained the mark scheme and trained the examiners in its use. The third group attended a coordination meeting that had a consensual format – there was an opportunity to discuss the mark scheme and, if the group’s views differed from the principal examiner, a consensus had to be reached and the mark scheme might be amended. The 36 examiners then all marked the same 45 scripts.

The researchers’ analysis showed that there was no measurable difference between the reliability of marking for the three different groups. But as with the previous study (by the same authors) described above, these examiners were experienced and already belonged to a community of practice.

The studies discussed here suggest that, for experienced examiners, a face-to-face meeting is not necessary to ensure reliable marking. However this does not give us any information about how reliability might be affected for new examiners who do not have the opportunity to meet the principal examiner and hear their explanation of the rationale for the mark scheme. With the advent of on-screen marking there are far fewer face-to-face coordination meetings for examiners. The process described in section 2.5 only includes meetings for the senior examiners, who are already part of the community of practice. Will new examiners become part of that community without the personal contacts that meetings bring? Of course there are many on-line communities in existence, so it is not impossible for a community to grow in this way.

3.2.4 Is reliability of marking affected by the type of mark schemes used?

The three categories of mark scheme used in GCSE examinations are objective mark schemes, points-based mark schemes and level-based mark schemes. The key features of each of these have been described in section 2.4.

Sütő and Nádas (2008) researched the marking of physics papers, all of the questions were objective or used a points-based scheme. They classified the marking strategy for each question as ‘simple’ or ‘complex’. For simple marking tasks it was straightforward to match a candidate’s answer to the correct answer given in the mark scheme. The mark schemes classed as complex were those where perhaps a key phrase was looked for or where the mark scheme required some evaluation of the quality of the response. They found a significant drop in accuracy in the marking of those questions where the marking task was more complex; the marker agreement for questions differed for apparently simple marking strategies (P_0 = 0.99, SD = 0.102) and apparently more complex marking strategies (P_0 = 0.78, SD = 0.14). This might be expected – simple marking strategies do not require qualitative decisions to be made; in my own experience, and that of colleagues, inaccuracies in marking straightforward objective questions are usually due lack of concentration. There are various reasons why answers with more complex marking strategies can lead to inaccuracies. The examiner may be required to make an evaluation of whether the answer given by the candidate matches the ‘spirit’ of the mark scheme when a candidate does not use exactly the same words as are given in the mark scheme. In other cases, the examiner may miss a marking point because there is poor handwriting or poor English (Meadows & Billington, 2005).

The findings of Sütő and Nádas (2008) are supported by other studies. In a study of inter-marker reliability for examiners marking non-live scripts (past papers) across five subjects, Massey and Raikes (2006) found a higher level of marker agreement for objective questions than for questions involving points-based marking and levels-based marking. For points-based marking they found that reliability decreased as the number of marks available increased; this might be expected – the examiner has to look for points within a longer piece of writing, and therefore is more likely to miss something, or reward a point that is not well enough made. Massey and Raikes found that, when they compared points-based and levels-based marking, there were differences between subjects, and that the relationship between the levels of agreement and number of marks was less clear. For an A-level Economics paper the inter-marker agreement was lower for the three-mark points-based questions (ICC_{mean} = 0.517) than for any of the three levels-based questions, each of which used three levels with maximum marks of 8, 10 and 12 (ICC_{mean} = 0.74, 0.567 and 0.585 respectively). An A-level Sociology paper used in the study consisted of 6 levels-based items, and all were marked with a good level of agreement (ICC_{mean} = 0.829). (For an explanation of ICC see section 3.2.2)

The studies of Massey and Raikes (2006) and Sütő and Nádas (2008) both compared the marking of non-live scripts by examiners – in this situation the pressure to be accurate is reduced compared to marking live papers where the results may affect candidates' futures.

In contrast, Bramley (2008) collected information about the marking of live papers. The research used scripts that had been marked by an assistant examiner and then remarked by the team leader as part of the monitoring process (see section 2.5). Bramley made a more detailed analysis of the factors that might affect marker agreement. Similar patterns were found as in other studies – in general, marker agreement decreased as the number of marks available for a question increased. There was an interesting comparison between marker agreement for points-based and levels-based marking. For questions with a tariff (the maximum mark for the question) between 5 and 9 the median value for the level of agreement between markers was similar for points based- marking and levels-based marking ($P_0 \approx 0.87$). For questions worth more marks (10-20) the agreement for levels-based marking was higher ($P_0 \approx 0.75$) compared with for points-based mark schemes ($P_0 \approx 0.55$).

In another study using data from the marking of live papers, Black, Sütő and Bramley (2011) looked at the level of marker agreement on seeding items, that is samples of work for which the mark has been agreed at the standardisation meeting (see section 2.5). As in Bramley's (2008) earlier work, they found that items with higher tariffs were associated with lower values of P_0. However in contrast to the earlier work, Black et al. reported much lower values of P_0 than Bramley for both level-based and points based schemes (Figure 8), and also reported that level-based schemes had even lower P_0 values than point-based schemes.

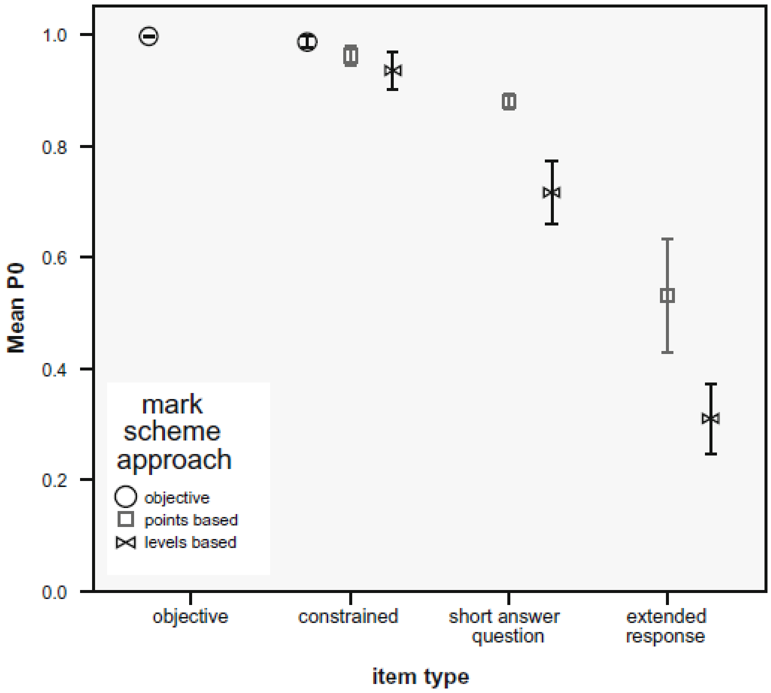


Figure 8 Marker agreement (mean P_0) by mark scheme approach. Error bars 95% (B. Black et al., 2011, p. 303)

The full data for the 2011 study of Black et al. are not available in the report, so it is not known which subjects were used, nor how many marks there were for each item, and whether the differences in P_0 varied with tariff.

In Bramley's 2008 study the value of P_0 was calculated using the amendments made by the team leader (TL) to the mark awarded by the assistant examiner for a script that the TL had not previously seen. As Bramley pointed out, a TL reviewing the marking of an answer which uses a best-fit approach to marking, may be more likely to tolerate a difference in mark which would not be tolerated on a points-based mark scheme where acceptable answers are more clearly defined. In contrast, in the 2011 study of Black et al. (2011) the mark for each seeding item was predetermined, so there was no room for tolerance between the examiner and the 'correct' mark for the script. This suggests that Bramley's proposal in 2007 that expecting examiners to agree to within one mark of each other for levels-based mark schemes might be more realistic than expecting a perfect match for mark schemes which require qualitative judgements.

3.2.5 Validity and reliability of questions and mark schemes

The validity of an assessment has a number of facets. Stobart (2009) has said that when describing the validity of an assessment, there are a number of factors that determine the overall validity, these include construct validity, fitness for purpose, and reliability. These terms are relevant to this study, Table 2 summarises the meaning he gave to these term.

Table 2 Extract from Table 1 A validity framework for national curriculum assessment (Stobart, 2009, p.165)

Concept	Inquiry	Potential threats to validity
Construct validity and fitness for purpose	What is being assessed?	Unclear construct; contested construct
	Does the assessment do what it is claiming to do?	Inadequate sampling of construct/domain (construct underrepresentation); Sampling of other constructs (construct irrelevance)
Reliability	How reliable is the assessment system?	Security breaches; inconsistent test administration and conditions; inappropriate modifications / time constraints; test-taker reliability
	How defensible are the results?	Inconsistent mark schemes; unreliable mark capture and aggregation; Insufficient data available for decision making; Inappropriate weightings; Level setting (grading processes inconsistent; Limited reference to previous standard setting

The outcomes from the studies described earlier indicate that an examination paper that uses objective questions and questions that can be marked using points-based mark schemes might yield more reliable marking. However there are some aspects of learning which do not lend themselves to these question styles. For high-tariff questions that demand an extended response, a levels-based mark scheme may be more appropriate. To not include questions of this type would reduce the construct validity of the assessment, that is, the assessment would not be able to assess all the knowledge, understanding, and aptitudes that the qualification is expected to reflect.

Thus there is a tension between ensuring the marking is reliable and producing a valid assessment. As Ofqual (2013c) states in its review of the quality of marking:

> …we must, therefore, accept that the exam system will never be able to deliver absolute reliability if we are to measure the right skills, knowledge and abilities in the right way. It does, however, need to be reliable enough to ensure that exam results can be used for their various high-stakes purposes, including accountability. (p. 21)

3.3 Argumentation in science education

In recent years there has been an increased interest in the role of argumentation in science teaching. This section reviews the literature that relates to the aspects of argumentation in science education considered in this study.

3.3.1 Using the terms ‘argumentation’ and ‘explanation’ in a science teaching context

Much of the research and development related to argumentation in the science education context draws on the work of Toulmin (1958, 2003). Toulmin wrote that the primary function of an argument is to support an assertion or claim. He identified a pattern of argument that is common across different fields, using the terms claim, data, warrant, backing and rebuttal. The relationships between these terms and the meanings he ascribed to them are shown in Figure 9.

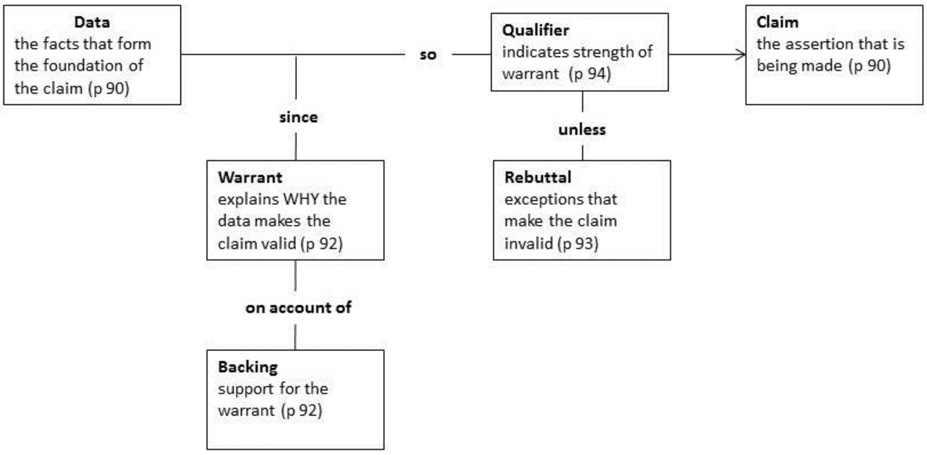

Data
the facts that form the foundation of the claim (p 90)
so
Qualifier
indicates strength of warrant (p 94)
Claim
the assertion that is being made (p 90)
since
unless
Warrant
explains WHY the data makes the claim valid (p 92)
Rebuttal
exceptions that make the claim invalid (p 93)
on account of
Backing
support for the warrant (p 92)

Figure 9 Toulmin's argument pattern (Toulmin, 2003)

Whilst this argument pattern can be used across many fields, what will count as acceptable backing and warrant for an argument will depend on what the argument is about. Figure 10 illustrates how Toulmin (2003) applied the ideas to an example from a legal context, seeking to establish that 'Harry is a British subject'. In this legal context the backing will often be reference to a particular statute, and the rebuttal is may refer to another statute that overrides the first.

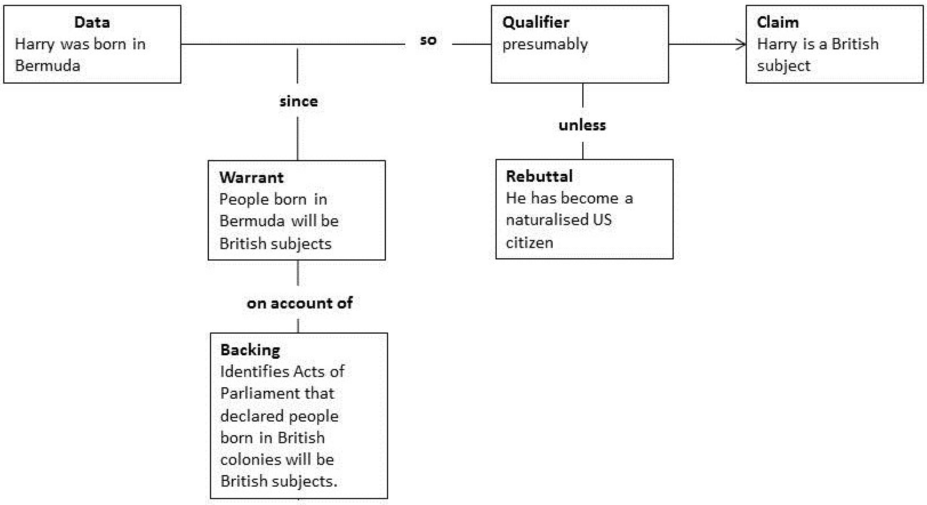

Data
Harry was born in Bermuda
so
Qualifier
presumably
Claim
Harry is a British subject
since
unless
Warrant
People born in Bermuda will be British subjects
Rebuttal
He has become a naturalised US citizen
on account of
Backing
Identifies Acts of Parliament that declared people born in British colonies will be British subjects.

Figure 10 An example of Toulmin's argument pattern (2003)

Osborne, Erduran and Simon (2004b) used Toulmin's argument pattern (TAP) in the Ideas, Evidence and Argument in Science (IDEAS) project in which they developed materials to train teachers in using argumentation in science lessons. Their interpretation of TAP for use in the field of science is shown in Figure 11.

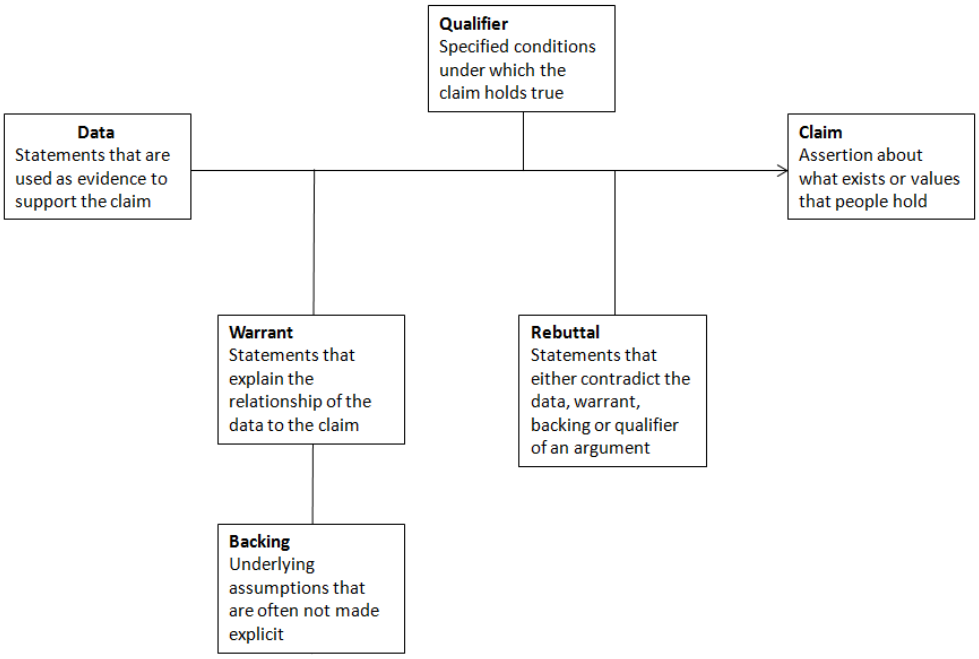
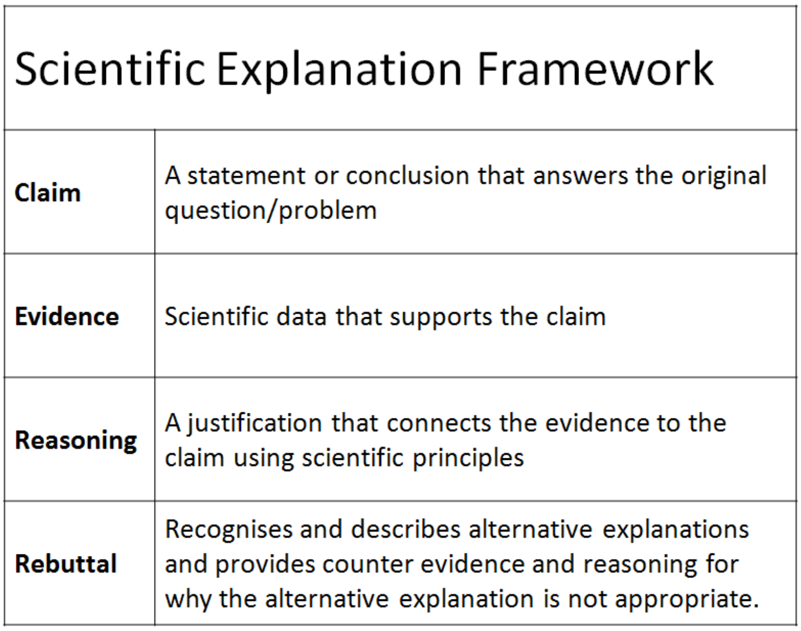

Qualifier
Specified conditions under which the claim holds true
Data
Statements that are used as evidence to support the claim
Claim
Assertion about what exists or values that people hold
Warrant
Statements that explain the relationship of the data to the claim
Rebuttal
Statements that either contradict the data, warrant, backing or qualifier of an argument
Backing
Underlying assumptions that are often not made explicit

Figure 11 TAP applied to argumentation in science (Osborne, Erduran & Simon, 2004b. pp 3.30-3.31)

There is an inconsistency in the training materials related to the use of the term rebuttal; Osborne et al. followed Toulmin's definition of rebuttal as shown in Figure 11, but in the guidance for trainers and teachers they also describe a rebuttal as the answer to counter arguments that someone else might make. This latter use of the word is closer to the way it is more commonly used in the literature related to argumentation in science (Chen, 2011; Khishfe, 2013; Simon & Richardson, 2009). In the IDEAS project pack the authors suggested a variety of topics that could be used to develop argumentation skills in science lessons, but they have not provided definitive 'answers' for teachers showing how specific arguments could be articulated using the TAP format. Perhaps this is partly because, as they acknowledged, it is sometimes difficult to distinguish between data, warrant, and backing (see also section 3.4). The materials for teachers used the formal language of argumentation, but Osborne et al. suggested that for discussion with students it might be more appropriate to use the terms 'claim', 'reasoning', 'grounds', 'justification', and 'evidence' and only use the term 'data', when referring to numerical values.

McNeill and Krajcik (2011) have also used some of the language of TAP in their work in the US to support teachers in teaching students to construct explanations in science. The terms they used and the meanings they assigned to them are shown in Figure 12.



Scientific Explanation Framework	
Claim	A statement or conclusion that answers the original question/problem
Evidence	Scientific data that supports the claim
Reasoning	A justification that connects the evidence to the claim using scientific principles
Rebuttal	Recognises and describes alternative explanations and provides counter evidence and reasoning for why the alternative explanation is not appropriate.

Figure 12 Scientific Explanation Framework (McNeill & Krajcik, 2011. p. xviii)

It can be seen that the terms are similar to those used by Osborne et al.(2004b), and they assigned similar meanings. Although McNeill and Krajcik used the same terminology as Osborne et al. they described the process as ‘scientific explanation’ rather than ‘argumentation’ and did not use the term argumentation anywhere in the resource. Elsewhere Krajcik and McNeill (2009) have cited the work of Toulmin (1958) and Osborne, Erduran & Simon (2004a), and others, but explained that “because our work focuses on classrooms, we chose to refer to this scientific practice as scientific explanation instead of argument to align with the language of the national standards” (p. 2).

Osborne and Patterson (2011) raised concerns that in the science education research literature the difference between an “argument” and “explanation” was becoming blurred and argued that it is important to maintain a distinction between the two constructs. They said that a science explanation “makes sense of a phenomenon based on other scientific facts” (p 629), whereas an argument uses evidence to justify a claim, for which there is a degree of tentativeness. This distinction is similar that made by Walton (1996), who wrote that the purpose of an argument is to settle an issue, to resolve some uncertainty attached to a proposition, whereas the purpose of an explanation is not to resolve whether some proposition is true, because that has already been accepted, rather it is to show why it is true. Osborne and Patterson (2011) argued that an important reason for distinguishing between explanations and arguments in teaching about science is that it will help students to come to an understanding of how we know what we know in science. Students should understand the explicit role of argumentation in the practice of science.

In response to Osborne and Patterson's 2011 paper, Berland and McNeill (2012) acknowledged the difference between the practices of argumentation and explanation but argued that it is not necessary to make the distinction when working with students in the classroom. They suggested that when students are asked to give scientific explanations in their work there is an overlap in their response between the explanation and the argument that justifies it and that emphasising the difference between argumentation and explanation may cause confusion about the scientific process rather than developing understanding.

In response to Berland and McNeill (2012), Osborne and Patterson (2012) argued that just because it may be difficult to make the distinction in the classroom does not mean it should not be done. They went on to argue that it is essential that the distinction between argument and explanation is made clear in official policy documents that guide the development of curriculum and assessment, so that teachers and examiners use the words appropriately. They point out that in the US (where Osborne, Patterson, Berland, and McNeill are all working) the National Research Council (2011) framework for next-generation standards makes the distinction; it suggests that by the end of grade 12 (age 18) "Increased sophistication, both of their model based explanations and the argumentation by which evidence and explanation are linked, is developed through mathematical and language skills appropriate to the grade level." (p. 239).

In England, the GCSE subject criteria for science (Ofqual, 2009) do not use the term argumentation but do use the terms argument and explanation; for example, "GCSE specifications in science must require learners to develop the ability to: …. develop arguments and explanations" (p. 6).   In the description of the work of Grade C candidates the document states that they "develop arguments with supporting explanations" (p.9). In Toulmin's model of argumentation these supporting explanations are referred to as warrants, so the ideas of argumentation are there in the criteria, but perhaps they are not very obvious to teachers and examiners.

3.3.2 Why teach argumentation in science lessons?

Those who advocate the teaching of argumentation in science lessons give a variety of reasons why they believe it is important. Three commonly cited reasons are that:

- argumentation is a key part of the practice of science. Therefore an essential part of any science curriculum must be understanding the role of evidence and argument in the scientific process – how we know what we know (Driver, Newton, & Osborne, 2000; Duschl & Osborne, 2002; Jiménez-Aleixandre, Bugallo Rodríguez, & Duschl, 2000; Kind, Kind, Hofstein, & Wilson, 2011; Osborne et al., 2004a; Simon & Richardson, 2009).
- students who develop argumentation skills will become more critical consumers of information, more ready and able to question the claims and arguments of others (Driver et al., 2000; Erduran & Jiménez-Aleixandre, 2008; Tiberghien, 2008).
- To convince a student why the scientific explanation is correct, they must also see why alternative explanations are wrong. In developing their own arguments to explain scientific ideas, students take ownership of the ideas and develop their conceptual understanding (Driver et al., 2000; Duschl & Osborne, 2002; Kelly, Druker, & Chen, 1998; Newton et al., 1999; Osborne, 2011).

The first two of these reasons for teaching about argumentation in science were made explicit in the Beyond 2000 report (Millar & Osborne, 1998) which aimed to address concerns that "The current curriculum retains its past, mid-twentieth-century emphasis, presenting science as a body of knowledge which is value-free, objective and detached – a succession of 'facts' to be learnt, with insufficient indication of any overarching coherence" (p. 8). The report recommended that all students should learn about the roles played by evidence and argument in establishing our knowledge and understanding of the natural world. The report authors suggested that by learning to make their own arguments and to evaluate the arguments of others they would develop skills that would be useful throughout their lives, both at work and in their personal lives.

The Twenty First Century Science Project (2006) drew on ideas in the Beyond 2000 report (Millar & Osborne, 1998). The project developed a new and flexible suite of courses for GCSE science, which are examined by the awarding body, OCR. The core science course in this suite is designed for all students, whether they intend to pursue a career in science or not, and it is intended to develop the scientific literacy of the students (Millar, 2006). The premise on which the core science course was developed was that an informed citizen needs both science content knowledge and also an understanding of the nature of science and of the ways in which science knowledge is obtained. The citizen needs to understand how such knowledge claims are tested through a process of argumentation. Within the Twenty First Century Science Project this understanding about the nature of science is made explicit in the Ideas about Science section in the course materials and in the specification (OCR, 2011).

3.4 Frameworks for analysing the quality of argument

As mentioned earlier, Osborne et al. (2004b) used Toulmin's Argumentation Pattern (TAP) (Toulmin, 1958) as the starting point for the development of the IDEAS project. They and other researchers who are interested in the quality of argument in science classrooms have also drawn on the same materials in developing frameworks for their research.

Kelly et al. (1998) used TAP as the starting point for an analysis of pairs of students' conversations about electric circuits. They focused on looking for whether the students used warrants to support their claims. They found that the circumstances under which students used warrants to back their assertions varied and was not necessarily a reflection of their ability in science. Students used warrants to convince their partner of their point of view but if both partners agreed on an answer there was no need for warrants, unless prompted by a teacher. The authors pointed out there are uncertainties in analysing a conversation and assigning labels to elements of the conversation. Students were not using the formal structure of the Toulmin pattern in their conversation. For instance there were occasions when a statement could have been described as a ‘claim’, or as a ‘warrant’.

Erduran, Simon and Osborne (2004) used TAP to analyse the quality of teacher-mediated arguments and also the quality of students’ discourse. Their aim was to design a system of analysis that was quantifiable and could subsequently be used to identify changes in the quality of argumentation, in order to measure the effectiveness of interventions.

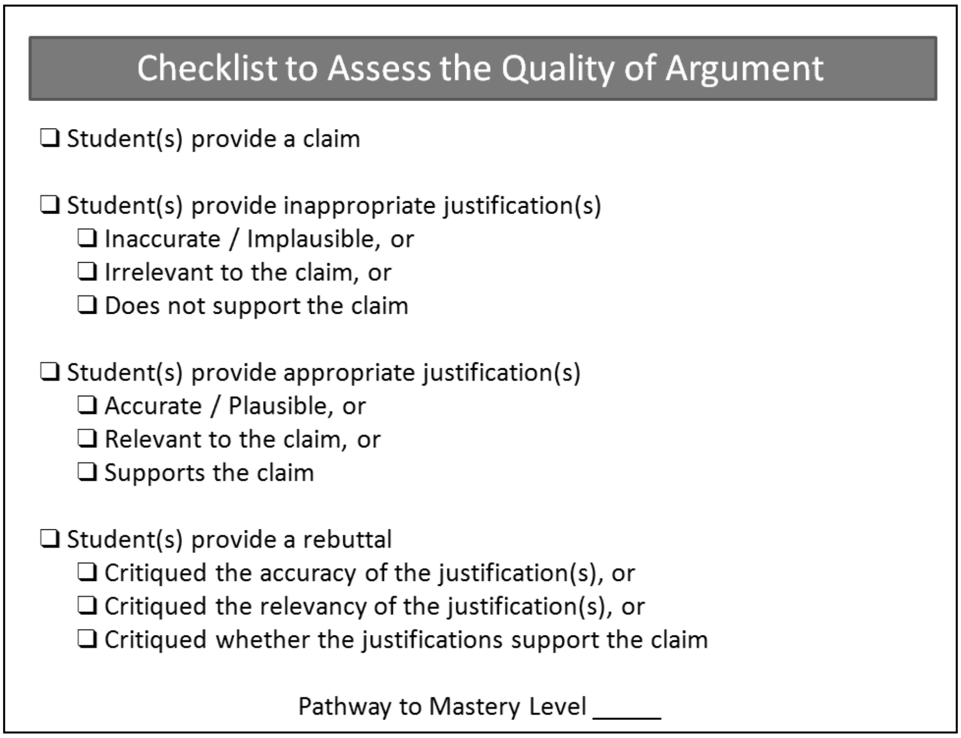
Like Kelly et al (1998), Erduran et al found some difficulty in mapping the conversations they were analysing to the Toulmin pattern. It was not easy to distinguish between data and warrants and between warrants and backing in analysing student discourse. They devised a framework that looked at whether the argument was supported by reasons (i.e. data, warrants or backing, without distinguishing between these). Secondly they looked for rebuttals of possible counter-arguments being included in the argument. They argued that, without a rebuttal, an argument does not challenge any counter-claims and that “the presence of rebuttals in conversation can act as an indicator of sustained engagement in argumentation discourse.” (p. 927). They assigned descriptions to five levels of argument (Table 3).

Table 3 Analytical framework used for assessing the quality of argumentation (Erduran et al., 2004. p. 928)

Level 1	Level 1 argumentation consists of arguments that are a simple claim versus a counter-claim or a claim versus a claim.
Level 2	Level 2 argumentation has arguments consisting of a claim versus a claim with data, warrants, or backings but do not contain any rebuttals.
Level 3	Level 3 argumentation has arguments with a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal.
Level 4	Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.
Level 5	Level 5 argumentation displays an extended argument with more than one rebuttal.

Kind et al. (2011) used the framework devised by Erduran et al. (2004) to determine the quality of the argumentation that took place between students whilst carrying out laboratory-based tasks. They found that most arguments only reached level 2, and that higher levels of argument generally took place when students were evaluating the results of their data.

Knight and Grymonpre (2013) used the ideas of McNeill and Krajcik (2011) (see section 3.3.1) to develop a framework to assess the quality of arguments of students aged 12-13 in the science classroom (Figure 13). Their framework can be used to assess written or spoken arguments and is intended to be shared with students so that they know the success criteria.



Checklist to Assess the Quality of Argument

❑ Student(s) provide a claim

❑ Student(s) provide inappropriate justification(s)
- ❑ Inaccurate / Implausible, or
- ❑ Irrelevant to the claim, or
- ❑ Does not support the claim

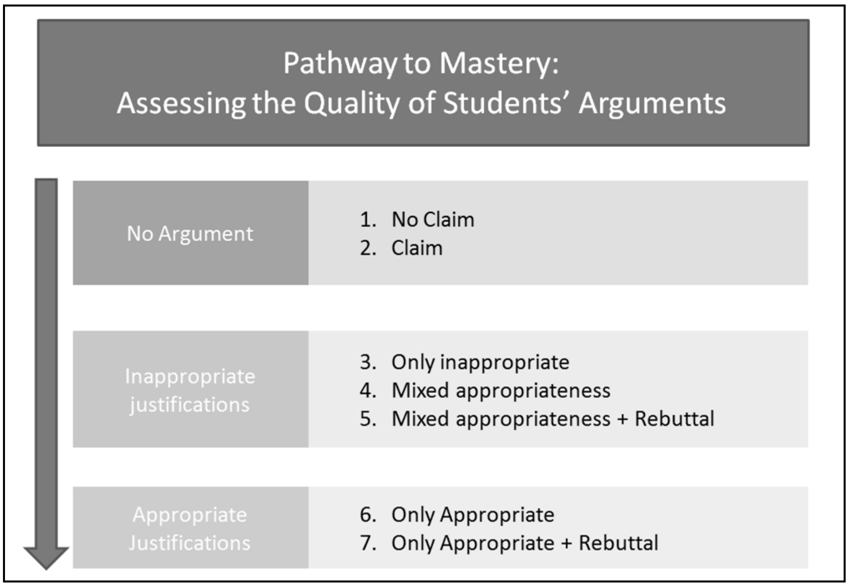
❑ Student(s) provide appropriate justification(s)
- ❑ Accurate / Plausible, or
- ❑ Relevant to the claim, or
- ❑ Supports the claim

❑ Student(s) provide a rebuttal
- ❑ Critiqued the accuracy of the justification(s), or
- ❑ Critiqued the relevancy of the justification(s), or
- ❑ Critiqued whether the justifications support the claim

Pathway to Mastery Level _____

Figure 13 Checklist to assess the quality of students' arguments (Knight & Grymonpre, 2013. p. 52)

The descriptions assigned to rebuttals in this framework do not match those used by McNeill and Krajcik (2011) or Osborne et al. (2004b). To make progression visible, Knight and Grymonpre assigned benchmarks to stages on the ‘pathway to mastery’ (Figure 14).



Pathway to Mastery:
Assessing the Quality of Students' Arguments

No Argument
1. No Claim
2. Claim

Inappropriate justifications
3. Only inappropriate
4. Mixed appropriateness
5. Mixed appropriateness + Rebuttal

Appropriate Justifications
6. Only Appropriate
7. Only Appropriate + Rebuttal

Figure 14 Pathway to mastery (Knight & Grymonpre, 2013. p. 52)

The frameworks developed by Erduran et al. (2004) and by Knight and Grymonpre (2013) were developed during work with middle-school students, but the descriptors are very general and hence they are applicable to students across a wider age range – the demand of the task and progression through the levels will depend upon the students' understanding of the science content as well as their mastery of argumentation skills.

3.5 Assessment of argumentation in GCSE science

As mentioned in section 1.3, the GCSE Science Criteria (Ofqual, 2009) stated that specifications should require learners to develop arguments and explanations. In the same document the grade description for an A grade candidate includes a requirement to "to develop arguments and explanations taking account of the limitations of the available evidence. They make reasoned judgments consistently and draw detailed, evidence-based conclusions." (page 8).

In a survey across the GCSE Science specifications from all the awarding bodies (AQA, 2011a, 2011b; Edexcel, 2013b; OCR, 2011, 2012c) the word 'argument' rarely occurs, apart from where it is part of the text about the aims of the specification and the grade descriptions, as required by the Ofqual criteria (2009) quoted above. The exceptions to this are the OCR specifications for Gateway Science (2012c) and Twenty First Century Science (2011).

The Gateway specification (OCR, 2012c) does suggest teaching opportunities for students to “develop the skills of scientific argument”, however in the assessed outcomes it only requires candidates to “identify arguments for and against scientific or technological development” (p3), but not develop their own arguments.

In the Twenty First Century Science specifications (OCR, 2011) the section ‘Making decisions about science and technology’ includes a requirement that candidates be able to “in a given context, identify, and develop, arguments” (p20). This suite of papers regularly includes questions that require candidates to use data to support their scientific explanations – in other words they are implicitly being asked to carry out an argument to support their assertions.

For example, part (a)(ii) of question 5 from the GCSE biology paper A161/1 in January 2012 (OCR, 2012d), which is shown in full in Figure 38 in Appendix 1, requires candidates to use data to support their conclusions. The question is about environmental indicators and asks candidates to comment on another student’s explanation of the data. The mark scheme states that a Level 3 answer “gives an explanation of how insecticide use in nearby fields could affect the river water and the species in the river and making (sic) appropriate references to the data.” (OCR, 2012e). The principal examiner’s report to centres reflects the expectation that candidates will use evidence to support arguments.

> 5(a)(ii) Those candidates who chose to disagree with the overall conclusions of the insecticide investigation struggled to identify supporting evidence. This restricted their score for this item. Using the data fully was credited fully and many students were able to achieve an acceptable level of response, supported by clear references to the values provided in the data tables. (OCR, 2012a)

Principal examiners reporting to centres make similar observations about candidates’ responses to other questions which ask for explanations (OCR, 2012b, 2012s). Although not made explicit in the question, examiners expect candidates’ answers to show the characteristics of a good argument: they should use evidence to support their assertions (claims), explain their reasoning (provide warrants and backing) and in some cases identify the counter arguments (rebuttals).

If this is what examiners are looking for, it could be argued that this should be made more explicit in the specification and in the examination questions. This issue will be returned to in chapter 8.

3.6 Summary

There is a consensus amongst many science educators that students should learn to use evidence to support their arguments; the GCSE Science criteria state that their ability to do this should be assessed in GCSE Science examinations. The requirements of such an assessment are not made clear in specifications, and it may be helpful to teachers and candidates if mark schemes made explicit what is expected when a questions requires students to support their argument with evidence.

A common framework based on a theoretical model of argumentation might provide the opportunity to develop mark schemes which had common features, which could help to provide consistent standards and reliable marking across subjects and between examination sessions. Such a framework would need to be shown to generate valid mark schemes.

Chapter 4 Methodology – research strategy and methods

4.1 Outline

This chapter describes the purpose of the study and identifies the research questions it aims to answer; it then outlines the stages of the work, and the rationale for the strategy and methods used.

4.2 The purpose of the study and the research questions

The purpose of this study was to find out if it would be useful to base the mark schemes that are used to assess extended answers in GCSE science on a generic framework which is informed by theoretical ideas about argumentation and by the expected demand of questions described by the grade descriptors. If such a framework could be developed it might help to ensure that mark schemes have a consistency of demand for questions targeted at the same grades on different papers and at different sessions (see section 2.4). The research questions that directed the study were:

> **Q1** Can a theoretically-based model of argumentation be used as the basis for developing a framework to evaluate extended answers to questions in GCSE Science examinations that require an argument or explanation?
>
> **Q2** How do the marks awarded when using a mark scheme based on such a framework compare with those awarded using the conventional mark scheme from the awarding body?
>
> **Q3** Would examiners find such a framework for writing mark schemes useful in establishing consistency of demand across different papers within the suite and year on year?

4.3 Research strategy

The research strategy adopted is outlined in Figure 15. The research question(s) that were addressed by each stage in the process are indicated in the boxes. Answering

Q2 and Q3 will be dependent on the answer to Q1 being ‘yes’, so development of the framework has to be the first stage in the research.

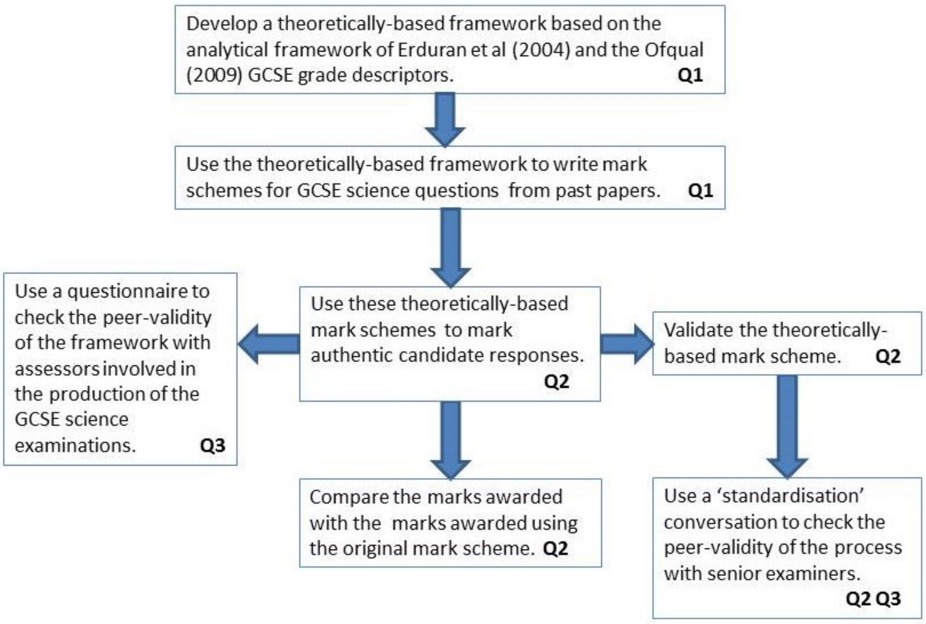

Develop a theoretically-based framework based on the analytical framework of Erduran et al (2004) and the Ofqual (2009) GCSE grade descriptors. Q1
Use the theoretically-based framework to write mark schemes for GCSE science questions from past papers. Q1
Use a questionnaire to check the peer-validity of the framework with assessors involved in the production of the GCSE science examinations. Q3
Use these theoretically-based mark schemes to mark authentic candidate responses. Q2
Validate the theoretically-based mark scheme. Q2
Compare the marks awarded with the marks awarded using the original mark scheme. Q2
Use a ‘standardisation’ conversation to check the peer-validity of the process with senior examiners. Q2 Q3

Figure 15 Research strategy for the study

Assuming that theoretically-based mark schemes could be developed, they need to be tested – this is the basis for Q2. This was addressed by using the schemes to mark answers given by candidates taking GCSE science examinations. The marks awarded by the researcher were compared with those awarded by the examiners who had marked the scripts using the original mark scheme from the awarding body.

To ensure that mark schemes based on the theoretical framework could be used by people other than the developer, the validity of the schemes developed in this study was checked by asking senior examiners to apply the schemes to the same sets of scripts; the marks awarded by the researcher and the senior examiners were compared.

The rationale for Q3 is that such a framework has no value unless it is useful to principal examiners setting and marking papers. The principles that guided the development and its role in developing mark schemes was described to those who write, revise and lead the marking of GCSE science papers and they were asked, via an open questionnaire, about the usefulness of the approach.

4.4 Methodology

4.4.1 Theoretical basis of the framework

The framework developed was intended to provide generic grade descriptors that could be used to write levels-based mark schemes (see section 2.4.2) for questions that require an explanation or an argument.

The theoretical basis for the study was the analytical framework devised by Erduran et al. (2004) (see section 3.4). This framework provided a hierarchy of statements describing quality of argumentation (Table 3). It has been 'tried and tested' by Erduran et al. and also by other researchers (Kind et al., 2011). The framework was originally developed to analyse discussions in class; in this study provided the starting point for developing a framework to analyse written arguments.

To be of practical value to examiners it would be helpful to demonstrate a visible relationship between the framework and the grade descriptors in the GCSE subject criteria for science (Ofqual, 2009). Therefore the framework developed in this study also took note of the Ofqual descriptors

4.4.2 Source of examination questions, mark schemes, and candidates' scripts

The theoretically-based framework developed is intended to provide a common basis for writing mark schemes for GCSE science questions. To test whether this was possible the framework was used to write mark schemes for a selection of questions taken from GCSE science examinations, for which there were already 'official' mark schemes available. The questions for this study were taken from examinations set for the OCR GCSE Science A Twenty First Century Science suite in January 2012 (see section 3.5).

The research is of interest to staff in the Standards Division of OCR, and they agreed to provide samples of candidates' work and their marks and also statistical information about all the questions on the papers from which these questions came. Not all this information is available in the public domain. To meet data protection requirements the samples of scripts were selected and anonymized by staff at OCR. To limit the work load for staff, only a limited number of scripts could be requested.

Using candidates' work from public examinations rather than collecting answers given by students specifically for this study had a number of advantages:

- Candidates answering questions as part of their GCSE examination are likely to take the task seriously and make a good attempt at the questions.
- The questions and original mark schemes have been developed using the quality assurance procedures of the awarding body.
- There is data available which describes how the whole cohort of candidates responded to the questions in the examination, which indicates whether the question did in fact address the target grades for which it was intended.
- The marks derived from the theoretically-based schemes can be compared to those awarded using the original mark schemes.

4.4.3 Choice of examination questions to be used in the study

For this research the examination questions used were selected from those that require students to write an explanation supported by data and reasoning or describe an argument that uses scientific reasoning. The questions identified were targeted at each of the grades A, B, C, and D (see section 2.4.1). This allowed the full range of the framework to be tested; questions targeting grades lower than D do not ask generally ask for arguments and explanations.

Information from item level data about how the questions discriminated between students across the range of abilities was used in making the selection (see section 2.7); if the question or original mark scheme did not discriminate well between candidates it may not be possible to discover whether the theoretically-based mark scheme is effective.

To ensure the full range of levels in the mark scheme can be tested, four scripts were requested at each of 2, 3, 4, 5, and 6 marks for each of the five questions identified. In the event, because OCR provided full scripts, there were 26 scripts available for two of the questions used in the study.

4.4.4 Methods of comparing the outcomes from using the two mark schemes

The analysis compared the marks awarded by the examiner using the original mark scheme with the marks awarded using the theoretically-based mark scheme. Whilst it would not be expected that the marks awarded by the two different mark schemes would be identical, a clear correlation would be expected. Both mark schemes are seeking to differentiate between the candidates who answer the question well and candidates who answer less well, so the 'better' scripts would be expected to get higher marks with both systems. The purpose of the study is not to show whether a theoretically-based mark scheme is better for a specific question, but to find out if the theoretically-based framework can be used to produce valid mark schemes across a range of questions. A perfect correlation for each question would show that the original mark schemes, developed from examiners' professional experience of the standards required, gave the same outcomes as the theoretically-based mark schemes. A very weak correlation would suggest that the two schemes were measuring very different things.

Bramley (2007) suggested that a correlation coefficient is not helpful as a comparator when looking at the agreement between markers, as there may be a high correlation even when one marker is more harsh or more generous than the other. However in this research it is helpful to know whether the two mark schemes yield similar rank orders of candidates, even if there is a difference in severity of marking. One of the initial selection criteria for the sample answers used in this research was to include equal numbers of questions at each mark from 2 to 6, yielding a stratified random sample. Therefore the scores used in the research did not have a normal distribution, and non-parametric tests provide a more appropriate approach to the statistical analysis. There are two possible non-parametric tests used to compare rank order, the Spearman rank order correlation coefficient (Spearman's rho coefficient, ρ) and the Kendall rank order correlation coefficient (Kendall's tau coefficient, τ) (Siegel, 1956). Kendall's tau coefficient is recommended where the data set is small and there are ties in position (Field, 2009; Robson, 2011). The data sets used here each had scores for between 19 and 26 candidates, and as each of the candidates' scores can only be in the range 0-6 there are certain to be ties in position. Kendall's tau coefficient was therefore used to compare the rank orders. The statistics package SPSS was used to calculate Kendall's tau coefficient.

Bramley (2007) proposed P_{agr1} as a measure of marker agreement between examiners using levels-based mark schemes, where P_{agr1} is the proportion of scripts for which the differences between examiners' marks are not more than ±1 (see section 3.2.2). Other indicators used by Bramley included the mean of the differences between markers and the mean of the absolute differences. These three indicators were used to compare marks in this study, where all the questions used had a tariff of 6. Microsoft Excel was used to calculate these indicators and generate charts.

4.4.5 Checking the validity of the mark scheme

This study uses the term validity in a similar way to Stobart (2009), described in section 3.2.5; validity encompasses both the construct validity (is the mark scheme set out in a way that will cause markers to reward the things that the examiner was looking for) and reliability (would a different examiner give the same marks for the same reasons).

In a live examination situation mark schemes are usually used by many examiners, most of whom will have not been involved in writing the scheme but will be expected to apply it reliably after practice and feedback (section 2.5). For this reason the mark schemes developed in this study were trialled in use by three experienced senior examiners and their marking was compared with that of the researcher. These examiners were all chief examiners responsible for the setting and marking of the questions used in the study.

The statistical methods used to make comparisons were the same as those described in section 4.4.4.

After the examiners had used the mark scheme on study questions, the researcher met with them, the format of the meeting was similar to the standardisation meeting that senior examiners would hold to agree on the final mark scheme for a live examination (see section 2.5). Each answer was discussed in turn to establish how the examiners had interpreted the mark scheme and the group came to an agreement about the mark the answer should be awarded (referred to as the standardised mark in Appendix 4 . The meeting was recorded.

4.4.6 Evaluating the usefulness of the framework to the examination process

To answer research question 3, the senior examiners who trialled the mark scheme to check its validity were also invited to discuss the general principles of the study. This conversation informed the development of a short self-completion questionnaire which was emailed to 25 assessors who are involved in setting and marking GCSE Science A examinations for OCR; the questionnaire was sent to the Chairs of Examiners, Chief Examiners, Principal Examiners, Revisers and Scrutineers (these roles are described in the Glossary). The questionnaire sought to discover the assessors' views on using levels-based mark schemes and whether they thought the theoretically-based framework would be useful in devising such mark schemes. The questions were left deliberately quite open in order not to lead the examiners to any particular answers.

Questionnaires that are answered by a large number of participants usually take a selected answer approach to enable a quantitative analysis. When the audience for the questionnaire is limited it becomes possible to use open questions, without the analysis becoming onerous (Robson, 2011). As the number of potential participants was limited a qualitative analysis was possible. The questionnaire was emailed to each participant with a letter explaining the purpose and some background information about the study. A qualitative analysis of the questionnaire was used to identify common themes.

Chapter 5 Developing and applying the framework

This chapter describes the development of the theoretically-based marking framework and its use to write mark schemes for five GCSE Science questions. Section 5.3 describes how the mark scheme for one of the questions was applied to candidates' responses to the question.

5.1 Development of the theoretically-based framework

The analytical framework devised by Erduran et al (2004) to analyse oral arguments between students in the science classroom (see section 3.4) consists of a hierarchical set of statements to describe the quality of argumentation observed (Table 3); this framework provided the theoretical basis for the marking framework devised in this study.

When questions are written for public examinations in the UK the target grade for the question is identified by the setter (see section 2.4.1). The GCSE subject criteria for science (Ofqual, 2009) describe the expected performance of candidates at Grades A, C and F. Table 4 shows a mapping of these Ofqual grade descriptors against the levels used by Erduran et al (2004). The reasoning for the decisions made is given in the third column.

Table 4 Mapping the Ofqual grade descriptors against the framework of Erduran et al. (2004)

Erduran et al. (2004) Framework	**Grade descriptors (Ofqual, 2009)**	**Reasoning for decisions made**
Level 5 argumentation displays an extended argument with more than one rebuttal		
Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.	Grade A: "Candidates recall, select and communicate precise knowledge and detailed understanding of science................. They evaluate information systematically to develop arguments and explanations taking account of the limitations of the available evidence. They make reasoned judgments consistently and draw detailed, evidence-based conclusions."	"take account of the limitations of the available evidence" considered as equivalent to giving an "identifiable rebuttal" in the sense Toulmin (2003) used "reasoned judgments consistently and draw detailed, evidence-based conclusions" suggests use of data with scientific 'warrants or backing'
	Grade B – no grade descriptor for Grade B	
Level 3 argumentation has arguments with a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal	Grade C: "Candidates recall, select and communicate secure knowledge and understanding of science............ They understand the limitations of evidence and develop arguments with supporting explanations. They draw conclusions consistent with the available evidence."	"develop arguments with supporting explanations." at this level is equivalent to "arguments with a series of claims ..... with either data, warrants, or backings" "understanding the limitations of evidence" in some contexts would be equivalent to a rebuttal
Level 2 argumentation has arguments consisting of a claim versus a claim with data, warrants, or backings but do not contain any rebuttals.	Grades D and E No grade descriptors for Grades D and E	The requirements of the Grade F descriptor are between the descriptions for of Level 2 and Level 1.
	Grade F: "Candidates recall, select and communicate their limited knowledge and understanding of science.......... Candidates interpret and evaluate some qualitative and quantitative data and information from a limited range of sources. They can draw elementary conclusions having collected limited evidence."	
Level 1 argumentation consists of arguments that are a simple claim versus a counter-claim or a claim versus a claim.		Level 1 is below the description for Grade F.

The levels-based mark schemes used to mark extended writing questions in GCSE Science are divided into three levels with a descriptor given for each level (see section 2.4.2), which means there is a need for more finely grained descriptors than those in the framework of Erduran et al. So the next stage in developing the marking framework was to write intermediate descriptors for each of the three levels that would be needed for levels-based mark schemes which might be targeted at any grade between A and D.

In this theoretically-based framework the level 3 description is written to be consistent with the Ofqual grade descriptor for the target grade of the question (see section 2.4.1). Table 5 shows the full range of descriptors from the top level descriptor (A3) for a question targeted at grade A down to the lowest descriptor for a question targeted at grade D (D1). In this framework, the three level descriptors A3, A2, and A1 would be used as the basis for a mark scheme with target grade A (see Appendix 1 Figure 25 for an example of this in practice). Similarly a scheme targeted at grade B would use the three descriptors B3, B2, and B1, and so on for other target grades. As can be seen, there is an overlap of descriptors; for example, the mid-level descriptor for a grade A answer (A2) is also the top-level descriptor for a grade B answer.

The demand of a question for a candidate is determined both by the requirements of what the candidate is asked to do (described by the grade descriptors) and also by the difficulty of the science content.

Table 5 Level descriptors developed from the Erduran et al (2004) framework and grade descriptors

Erduran et al. (2004)Framework	**Grade descriptors**	**Level descriptors**
Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.	**Grade A**: "Candidates recall, select and communicate precise knowledge and detailed understanding of science................ They evaluate information systematically to develop arguments and explanations taking account of the limitations of the available evidence. They make reasoned judgments consistently and draw detailed, evidence-based conclusions."	**A3** The argument or explanation of a claim is supported by evidence (data) with clear scientific reasoning (warrant and backing). The argument takes account of the limitations of the evidence or provides a rebuttal to possible counterarguments. No serious errors of science.
Level 3 argumentation has arguments with a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal	**Grade B** – no grade descriptor for Grade B	**A2 B3** The argument or explanation of a claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence/argument.
	Grade C: "Candidates recall, select and communicate secure knowledge and understanding of science........... They understand the limitations of evidence and develop arguments with supporting explanations. They draw conclusions consistent with the available evidence."	**A1 B2 C3** The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.
Level 2 argumentation has arguments consisting of a claim versus a claim with data, warrants, or backings but do not contain any rebuttals.	**Grade D** No grade descriptors for Grades D	**B1 C2 D3** The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning (warrant) OR refers to limits of evidence.
		C1 D2 May make clear the claim; Provides some relevant evidence or scientific reasoning. No reference to limitations of evidence or reasoning.
	Grade F: "Candidates recall, select and communicate their limited knowledge and understanding of science.......... Candidates interpret and evaluate some qualitative and quantitative data and information from a limited range of sources. They can draw elementary conclusions having collected limited evidence."	**D1** Identifies some relevant factor, evidence or reasoning but the links are weak

5.2 Using the framework to develop mark schemes for specific questions

In order to find if this theoretically-based framework can be used in practice, and to address research question 1 (section 4.2), mark schemes were developed for each of the five examination questions selected for the study. Table 6 lists the questions that were used. The texts of the questions and both mark schemes (the OCR mark scheme and the mark scheme developed in this study) are included in Appendix 1.

Table 6 Questions from OCR Science A January 2012 used in the study

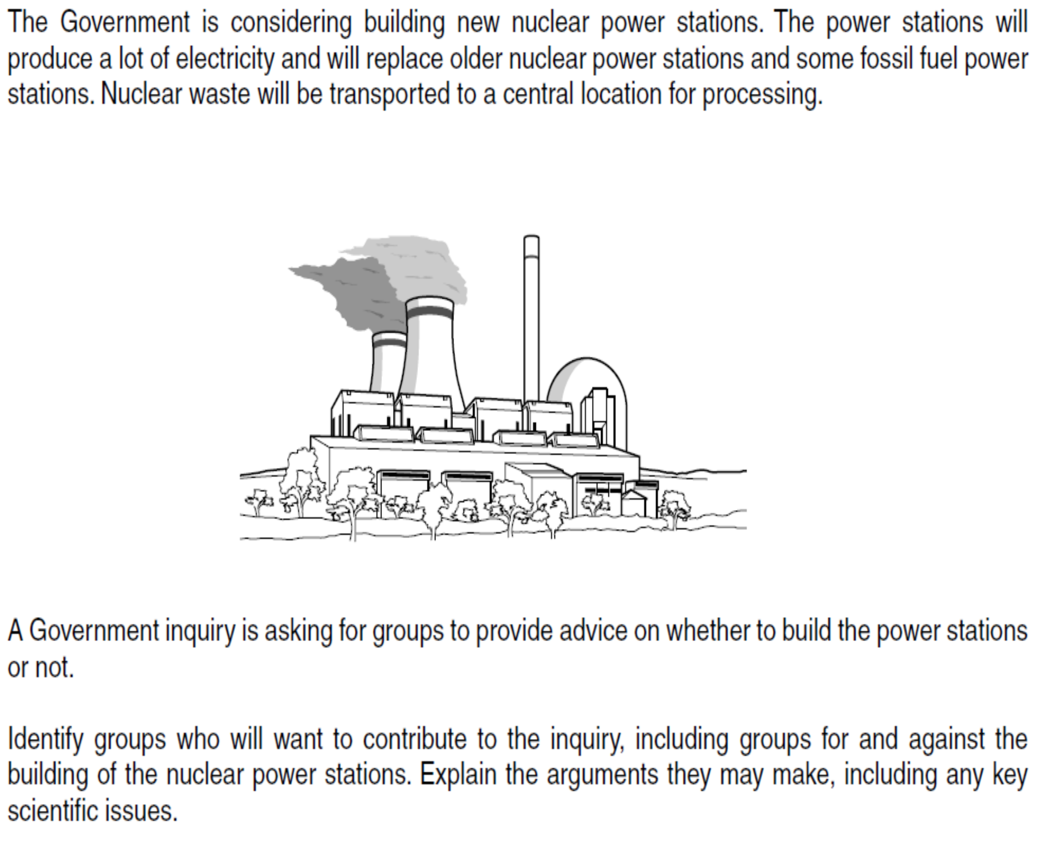
Question ID	Paper and question number	Topic of question	Level of demand
1	GCSE Chemistry A171/02 Q3ci (OCR, 2012i)	Properties of polymers	A
2	GCSE Physics A181/02 Q6 (OCR, 2012r)	Siting a nuclear power station	B
3	GCSE Chemistry A171/01 Q2b (OCR, 2012h) and A171/02 Q1b (OCR, 2012i)	Particulates and asthma*	C
4	GCSE Physics A181/01 Q6 (OCR, 2012q) and A181/02 Q4 (OCR, 2012r)	Risks of sunbathing*	C
5	GCSE Physics A181/01 Q9(OCR, 2012q)	Siting of a power station	D

*These questions were included in both the higher tier and foundation tier papers.

For each question, the theoretically-based framework was used to write a levels-based mark scheme. GCSE Science levels-based mark schemes are required to include descriptors for the quality of written communication in answers. The same descriptors are used in all OCR GCSE Science examinations, and these are included in the study mark schemes. Apart from this the mark schemes were developed without reference to the original OCR mark scheme used by examiners. Each mark scheme was developed by interpreting the general descriptors at each level to describe what would be required for the specific question.

Here the process is described in detail for study question 2 (see Appendix 1). The theoretically-based mark schemes for all 5 study questions are shown in Appendix 1.

Study question 2 (Figure 16) is from the GCSE Physics higher tier paper, (OCR, 2012r). This question addresses the part of the specification which is about generation of electricity and which specifically includes reference to nuclear power, nuclear waste and the effects of ionising radiation.



The Government is considering building new nuclear power stations. The power stations will produce a lot of electricity and will replace older nuclear power stations and some fossil fuel power stations. Nuclear waste will be transported to a central location for processing.

A Government inquiry is asking for groups to provide advice on whether to build the power stations or not.

Identify groups who will want to contribute to the inquiry, including groups for and against the building of the nuclear power stations. Explain the arguments they may make, including any key scientific issues.

Figure 16 Study question 2 A181/02 Q6 (OCR, 2012d)

The target grade for the question is grade B so the mark scheme was developed by interpreting the general descriptors at levels B3, B2, and B1, to describe what would be required for this specific question. The B3 description states that the argument should include rebuttals, but in this case candidates were asked to describe the arguments for each side so it was thought unreasonable to expect them also to explicitly provide the rebuttals that each side would give. The B2 descriptor refers to ‘limitations of evidence’ but this is not considered applicable to this context. The first draft of the mark scheme is shown in Table 7. The guidance column gives examples of the sorts of answers expected from candidates at a Grade B standard.

Table 7 Generic marking framework level descriptors and first draft of specific descriptors for study question 2

Framework descriptor	**Mark scheme** **Level of response descriptors**	**Guidance**
B3 The argument or explanation of the claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence or argument (weak rebuttal).	**Level 3** Answer identifies group for and puts forward at least one piece of evidence for with scientific explanation. Answer identifies group against and puts forward at least one piece of evidence against with scientific explanation. Quality of written communication does not impede communication of the science at this level.	**This question is targeted at grades up to B.** Answers at Level 3 must include reference radioactive materials / ionising radiation. (As candidates are required to put both sides of the argument, they are not expected to include any explicit rebuttals.) **Possible answers may include:** **Groups and arguments for nuclear power station** environmental groups – reduces CO_2 emissions of power production – so reducing greenhouse gases; reduces particulate/acid rain gases – so reducing environmental damage local people near old coal stations – less emissions from NPS so cleaner air; nuclear fuel much less bulky, so fewer lorries/rail trucks in and out workers near PS – provides work during demolition / construction of PS **Groups and arguments against power station** People living near NPS sites / People near nuclear waste disposal – concerns about ionising radiation during use / risk of accidents – radiation can cause cancer environmental group – disposal of nuclear waste is a problem: ionising, long lasting
B2 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence.	**Level 2** Answer identifies groups for and against nuclear power stations; uses evidence for and against with some scientific reasoning for at least one argument. Quality of written communication partly impedes communication of the science at this level.	
B1 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning, (warrant) OR refers to limits of evidence.	**Level 1** Answer identifies groups for and against nuclear power stations. Puts forward evidence for and against. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

This first draft of the mark scheme was tested by using it to mark a small number of scripts, as is the practice before the marking standardisation meeting (see section 2.5). Testing the mark scheme revealed that the level 1 descriptor in Table 7 was too demanding – there were candidates who described advantages and disadvantages of nuclear power stations without explicitly naming groups that would make those arguments. These answers certainly included relevant science and were worthy of credit, so were better than the Level 0 descriptor. The amended mark scheme is shown in Table 8 with the changed descriptor highlighted.

Table 8 Marking framework descriptors with amended descriptor (highlighted) for study question 2

Framework descriptor	**Mark scheme** **Level of response descriptors**	**Guidance**
B3 The argument or explanation of the claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence or argument (weak rebuttal).	**Level 3** Answer identifies group for and puts forward at least one piece of evidence for with scientific explanation. Answer identifies group against and puts forward at least one piece of evidence against with scientific explanation. Quality of written communication does not impede communication of the science at this level.	**This question is targeted at grades up to B.** Answers at Level 3 must include reference radioactive materials / ionising radiation. (As candidates are required to put both sides of the argument, they are not expected to include any explicit rebuttals.) **Possible answers may include:** Groups and arguments for nuclear power station environmental groups – reduces CO_2 emissions of power production – so reducing greenhouse gases; reduces particulate/acid rain gases – so reducing environmental damage local people near old coal stations – less emissions from NPS so cleaner air; nuclear fuel much less bulky, so fewer lorries/rail trucks in and out workers near PS – provides work during demolition / construction of PS Groups and arguments against power station People living near NPS sites / People near nuclear waste disposal – concerns about ionising radiation during use / risk of accidents – radiation can cause cancer environmental group – disposal of nuclear waste is a problem: ionising, long lasting
B2 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence.	**Level 2** Answer identifies groups for and against nuclear power stations; uses evidence for and against with some scientific reasoning for at least one argument. Quality of written communication partly impedes communication of the science at this level.	
B1 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning, (warrant) OR refers to limits of evidence.	**Level 1** Puts forward evidence for and against but may make not explicit links to groups. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

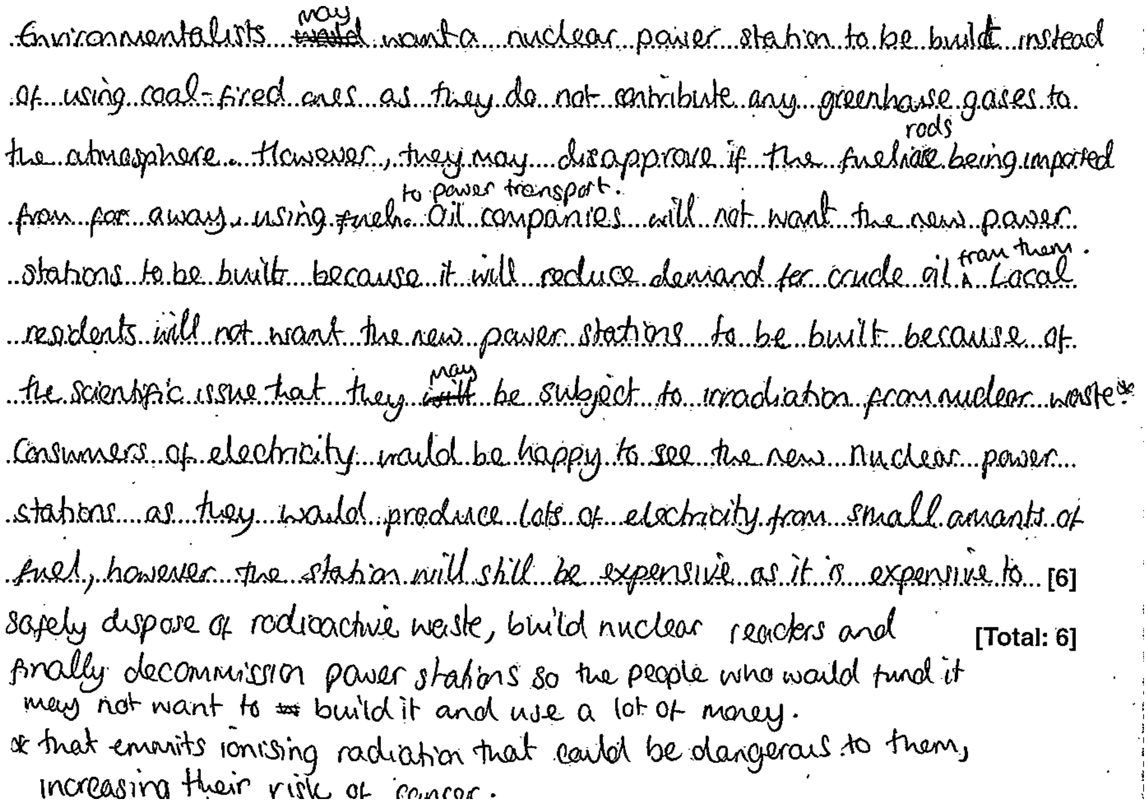
The theoretically-based framework in Table 5 was used to write mark schemes for all the study questions in the same way. The final theoretically-based mark schemes are shown in Appendix 1.

5.3 Applying the theoretically-based mark schemes

OCR supplied digital copies of candidates' scripts for each of the questions used in the study; the scripts had originally been marked on-line so there were no annotations or marks on them. The details of the marks awarded for each answer were supplied separately. The scripts were supplied in rank order of marks given for the study questions. So, to avoid any bias in marking arising from knowledge of the rank order, the answers used in the study were copied from the scripts and pasted into a new document in random order. They were then marked using the theoretically-based mark schemes.

When using a levels-based mark scheme the mark awarded is determined by first identifying which level descriptor best matches the answer. Whether the answer scores the higher or lower of the two marks is determined by how good the fit is. A partial fit will be awarded the lower mark; a good fit will be awarded the higher mark.

Figures 18, 19, and 20 are examples of three candidates' responses to study question 2. The text below each figure gives the rationale for the mark awarded for that answer using the mark scheme in Table 8.



Environmentalists ~~would~~ may want a nuclear power station to be built instead of using coal-fired ones as they do not contribute any greenhouse gases to the atmosphere. However, they may disapprove if the fuel rods are being imported from far away, using ~~fuel~~ to power transport. Oil companies will not want the new power stations to be built because it will reduce demand for crude oil from them. Local residents will not want the new power stations to be built because of the scientific issue that they ~~will~~ may be subject to irradiation from nuclear waste* Consumers of electricity would be happy to see the new nuclear power stations as they would produce lots of electricity from small amounts of fuel, however the station will still be expensive as it is expensive to safely dispose of radioactive waste, build nuclear reactors and finally decommission power stations so the people who would fund it may not want to ~~ss~~ build it and use a lot of money.

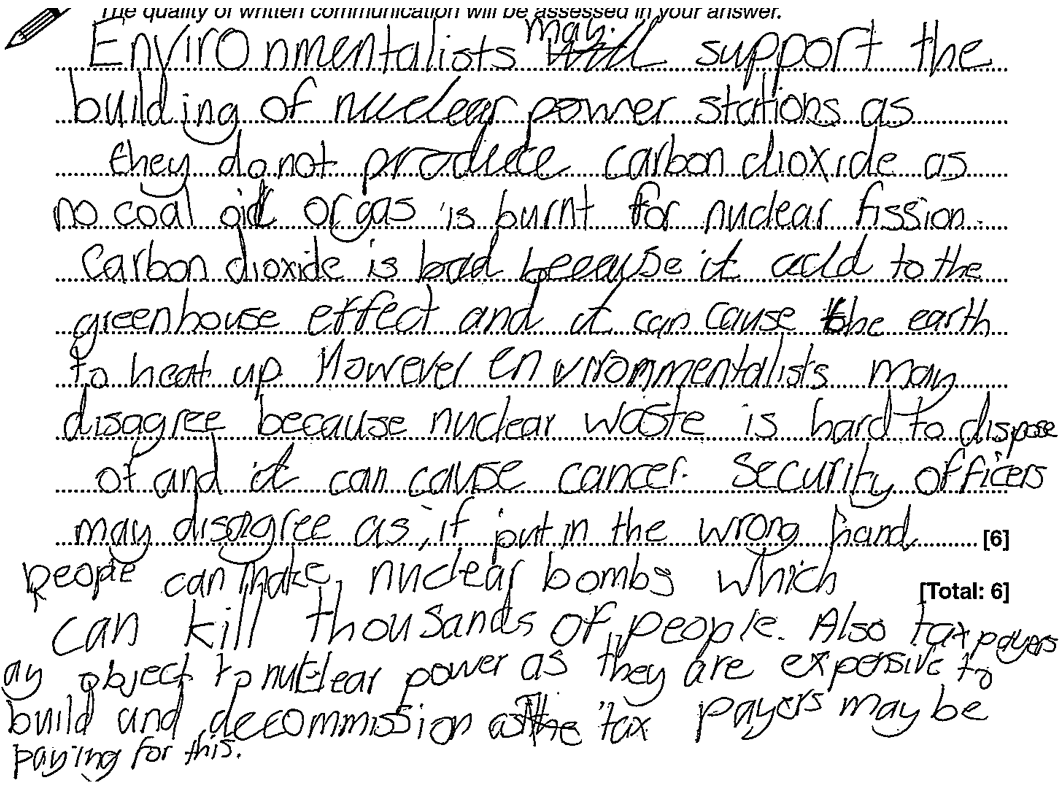
* that emmits ionising radiation that could be dangerous to them, increasing their risk of cancer.

[6]

[Total: 6]

Figure 17 Script O candidate code 1810218 (supplied by OCR) response to study question 2

In script O (Figure 17) the candidate identifies particular groups both for nuclear power stations being built (environmentalists) and against (oil companies and local residents). He/she gives scientific reasons for and against, although there is no scientific explanation of the significance of greenhouse gases. There is a developed reason against building a power station related to concerns about nuclear waste. This fits the level 3 description, but the argument for nuclear power lacks sufficient scientific detail and a mark of 5 was awarded.



The quality of written communication will be assessed in your answer.

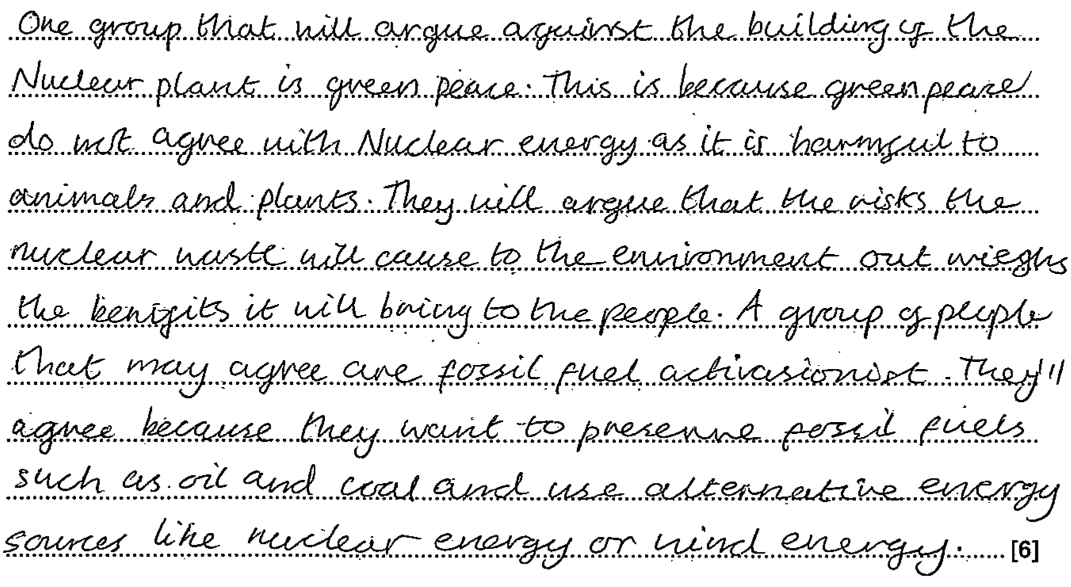
Environmentalists ~~will~~ may support the building of nuclear power stations as they do not produce carbon dioxide as no coal oil or gas is burnt for nuclear fission. Carbon dioxide is bad because it add to the greenhouse effect and it can cause ~~the~~ the earth to heat up However environmentalists may disagree because nuclear waste is hard to dispose of and it can cause cancer. Security officers may disagree as, if put in the wrong hand people can make nuclear bombs which can kill thousands of people. Also tax payers may object to nuclear power as they are expensive to build and decommission ~~as the~~ tax payers may be paying for this.

[6]

[Total: 6]

Figure 18 Script E candidate code 1810208 (supplied by OCR) response to study question 2

In script E (Figure 18) the candidate identifies groups for (environmentalists) and against (taxpayers, security officers and some environmentalists) building nuclear power stations. There is well developed reasoning about greenhouse gases and also the answer identifies the risks from nuclear waste. This takes it close to a level 3 answer. However the level descriptor for 3 states that there must be reference to radioactive materials / ionising radiation and neither of these is mentioned. So the answer best fits level 2 and 4 marks were awarded.



One group that will argue against the building of the Nuclear plant is green peace. This is because green peace do not agree with Nuclear energy as it is harmful to animals and plants. They will argue that the risks the nuclear waste will cause to the environment out wieghs the benifits it will bring to the people. A group of people that may agree are fossil fuel activasionist. They'll agree because they want to preserve fossil fuels such as oil and coal and use alternative energy sources like nuclear energy or wind energy. [6]

Figure 19 Script L candidate code 1810220 (supplied by OCR) response to study question 2

In script L (Figure 19) the candidate identifies a group against (Greenpeace) but does not give any scientific reasoning, simply mentioning 'nuclear waste' without any explanation. The naming of 'fossil fuel activists' is not an acceptable answer for a group of people in favour of nuclear power and although he/she does recognise that nuclear power would reduce the use of fossil fuels, there is no explanation of why that might be an advantage. The answer does not meet the criteria for level 2. It fits the level 1 descriptor best and is awarded 2 marks.

Chapter 6 Findings - Quantitative analysis to compare mark schemes

6.1 Overview

This chapter begins by considering the facility for all the questions of with levels-based mark schemes across the papers set at the January 2012 session; this will provide the context for comments on the facility of each of the study questions in later sections. Section 6.3 describes in detail the statistical analysis of the marks awarded for study question 2 using the original mark scheme and the study mark scheme. The following section provides a summary of the comparisons between marks awarded using the two mark schemes made for all five questions used in this study. Section 6.5 describes outcomes of the quality assurance procedure that was used to check the validity of the mark scheme. The data for all the study questions can be found in Appendix 3.

6.2 Facility of questions that used levels-based mark schemes

The five questions used in this study are a subset of the 25 questions with levels-based mark schemes in the January 2012 series of papers for the OCR GCSE Science A. Table 15 in Appendix 2 shows data about the facility of each of those 25 questions.

In Table 15 the target grade for each question, which is stated in the mark scheme, is identified alongside the facility of each question at that target grade. Facility of a question at a grade is calculated using the mean mark for that question for candidates who achieved that grade on the paper (Elliott & Johnson, 2007). The mean facility at grade across all questions (0.48) shows that students who achieved the grade for which the question was targeted scored, on average, just under half marks for the question. This facility may seem low; it might be thought that students who achieved a particular grade overall should be achieving a level 3 answer and scoring 5 or 6 marks rather than fewer than 3 marks. Although overall candidates will pick up most of their marks for questions set at or below their final grade, they will not be expected to score every mark at the target grade. The overall facility for these questions may be lower than expected because these were the first set of science papers to include this style of question; it might be expected that as teachers get used to the demands of this question type, and students have more examples to practice with, the facility will increase.

It can be seen from Table 15 that there is a general trend for the facility at the target grade to decrease for lower target grades. This is in line with the experience of the researcher; weaker candidates on any paper tend to pick up marks in a seemingly random way across the paper, rather than getting all their marks from the questions targeted at their level and below. This tendency is perhaps also reflected in the fact that the mean facility at target grade is lower on the foundation tier papers (0.41) than on the higher tier papers (0.54). The very low facilities of some of the questions may also suggest that the assessors were over-optimistic in their target grades for these questions.

In Table 15 those questions that were targeted at grade C and that were included in both the foundation tier and higher tier papers are identified by # in the Question column. For all these questions the facility was greater on the foundation tier (mean = 0.54, S.D = 0.12) than on the higher tier paper (mean = 0.45, S.D = 0.11). Although the question and mark schemes for these pairs of questions were identical, they were marked by different teams of examiners. It might be inferred that the difference in facility was due to the way the mark scheme was applied. There are, however, other possible explanations; it could be that the candidates achieving a grade C on the higher tier paper were, on average, weaker than those who scored a grade C on the foundation tier paper (it is conceivable that some of those who scored a C on the foundation tier paper might have scored a B if they had taken the higher tier paper). It would be interesting to check whether this discrepancy in facilities for common questions is seen more commonly, across other specifications and other sessions.

6.3 Comparison of marks awarded by the two schemes for study question 2

The marking of study question 2 was described in Section 5.2. All the data related to this is shown in Appendix 3. The key data is also show in this section of the report.

Table 9 shows the item-level data for the question. The question had been targeted by the Principal Examiners to discriminate at grade B. The facility at grade B was 0.55, which is very close to the average facility at target grade for all levels-based questions on the higher tier papers (0.54). Other item-level data for this question is provided in Appendix 3.2 .

Table 9 Examination statistics for study question 2 for the whole cohort (CA, 2012j)

17163 candidates	**facility (cohort)**	**facility at each grade**								
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
A181/2 Q6	0.54	0.81	0.66	0.55	0.40	0.24	0.14	–	–	0.07

Table 10 shows the marks awarded for the scripts used in the study by the original examiner (EM) and the researcher using the theoretically-based mark scheme (RM).

Table 10 Marks awarded by the original examiner (EM) and the researcher (RM) for research question 2

Question 2 **Nuclear power station**		**mark awarded using original mark scheme**	**mark awarded for question using mark scheme derived from theoretical framework**	**difference between marks awarded**
Candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM – EM**
1810201	B	2	2	0
1810202	D	2	2	0
1810203	F	2	2	0
1810204	H	2	2	0
1810205	A	3	2	–1
1810206	C	3	3	0
1810207	J	3	2	–1
1810208	E	3	4	1
1810209	T	4	4	0
1810210	G	4	3	–1
1810211	S	4	3	–1
1810212	P	4	3	–1
1810213	I	5	3	–2
1810214	K	5	5	0
1810215	R	5	5	0
1810216	M	5	5	0
1810217	Q	6	5	–1
1810218	O	6	6	0
1810219	N	6	4	–2
1810220*	L	6	2	–4

*The data for Candidate 1810220 (script code L) was not included in the statistical analysis, as on inspection, 6 marks should not have been awarded to candidate L on the original mark scheme. The script for Candidate L is shown in Figure 19.

The distribution of the differences between the marks awarded by the original examiner and the researcher |RM–EM| is shown in Figure 20.

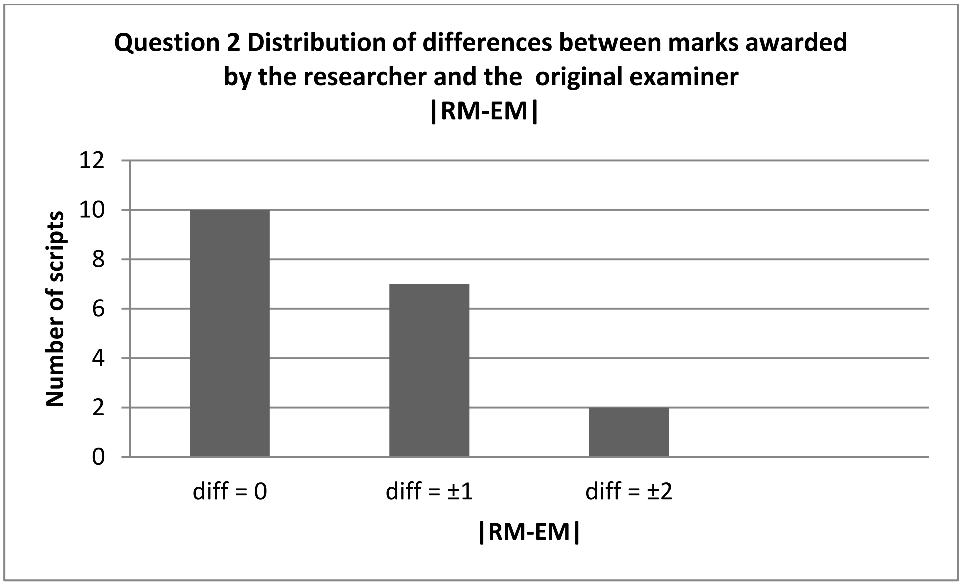


Figure 20 Study Q2 Distribution of differences between marks awarded by the examiner and the researcher |RM–EM|

The level of agreement between the two sets of marks $P_{Agr1} = 0.89$, that is 89% of the marks were within agreement ±1. The mean difference between the original examiner marks and the researcher marks RM–EM = – 0.53, signifies that the marks awarded using the study mark scheme were on average lower by about half a mark; the mean of the absolute differences (0.63) is less than 1, indicating that those marks that were not within ±1 did not differ hugely, (in this case, there were just two answers for which RM–EM was not within ±1, for both RM–EM = –2).

Kendall's tau coefficient $\tau = 0.751$ ($p< 0.001$) indicates a strong correlation between the rank orders given by the two mark schemes that is, the differences in marks awarded by the two schemes did not significantly affect the rank orders of the candidates.

6.4 Comparison data for all questions used in the study

The analysis described in section 6.3 was carried out for all the questions in the study, this is summarised in Table 11.

Table 11 Comparison of marks awarded by the two mark schemes

Study question	target grade	number of scripts[*]	facility at target grade	P_{Agr1}	Mean (RM–EM)	Mean absolute difference |RM–EM|	Kendall's tau correlation between EM and RM
Q1 Polymers A171/01 Q3ci	A	26	0.42	0.88	–0.31	0.54	0.827
Q2 Nuclear power station A181/02 Q6	B	19	0.55	0.89	–0.53	0.63	0.751
Q3 Asthma A171/01 2b	C	19	0.35	0.79	–0.53	0.74	0.614
Q3 Asthma A171/02 1b	C	26	0.27	0.85	0.00	0.71	0.632
Q4 Sunbathing A181/01 Q6	C	20	0.65	0.95	–0.25	0.45	0.824
Q4 Sunbathing A181/02 Q4	C	26	0.57	1.00	0.00	0.42	0.784
Q5 Power station A181/01 Q9	D	20	0.58	0.85	–0.65	0.75	0.744
mean				0.89	–0.32		
standard deviation				0.07			

[*]For some questions more than 20 scripts were available, and it was decided to include all scripts available. Only 19 scripts are included in the analysis for study Q2 (see section 5.3). Only 19 scripts were available for the foundation tier version of study question 3.

The analysis summarised in Table 11 suggest that the agreement between the two sets of mark schemes was generally good (mean $P_{Agr1} = 0.89$, SD = 0.07), that is in most cases the outcomes using the two schemes were within one mark of each other.

Overall the study mark schemes yielded slightly lower marks than the original mark schemes, with all mean differences ≤ 0, implying that the theoretically-based mark schemes were perhaps more demanding and/or applied in a more stringent way.

There were also good correlations between the rank orders generated by the two sets of mark schemes ($\tau = 0.614$ to 0.824). The weakest correlation was for study question 3. A comparison of the study mark scheme and the original mark scheme

(Appendix 1) shows there are some differences in the requirements of the two schemes. If the two schemes are not looking for the same characteristics in the answers it might be expected that there would be a weaker correlation between the rank orders of the candidates.

6.5 Assurance checks

The mark schemes written for live examination papers are frequently used by many markers, so if the theoretically-based mark schemes developed here are to be useful, they must be understood and be able to be applied by other markers. If other examiners are able to use the mark schemes and award marks comparable to those awarded by the researcher, this suggests that the mark scheme has validity – that is it is operating in the way intended (has construct validity) and is reliable (yields similar marks no matter who marks the script). To find out if this is the case, principal examiners were asked to use the mark schemes to mark study questions 1 and 2. All the data that resulted from this process is given in full in Appendix 4.

6.5.1 Validation of the mark scheme for study question 1

Study question 1 is taken from the GCSE Chemistry paper (see Appendix 1). One of the principal examiners for chemistry was asked to use the theoretically-based mark scheme to mark the same sample of candidates' answers as the researcher had marked. The principal examiner (PE) marked the scripts without any discussion with the researcher or feedback during the marking.

All the data from this work is shown in appendix 4, with the key data also displayed here. Table 12 shows the marks awarded by the researcher (RM) and by the principal examiner (PM) and the differences between them (RM–EM). The distribution of the differences is show in Figure 21.

Table 12 Marks awarded by the researcher (RM) and the principal examiner (PM) for study question 1 during the validation process

Question 1 Polymers		mark awarded for question using mark scheme derived from theoretical framework		differences in marks awarded
candidate code	script code	researcher mark (RM)	PE mark (PM)	RM–PM
1710201	B	2	2	0
1710202	Y	4	5	–1
1710203	U	0	0	0
1710204	H	6	6	0
1710205	A	2	4	–2
1710206	C	3	2	1
1710207	J	3	2	1
1710208	G	4	4	0
1710209	T	4	3	1
1710210	V	0	0	0
1710211	S	5	6	–1
1710212	P	1	0	1
1710213	W	0	0	0
1710214	K	0	0	0
1710215	R	2	0	2
1710216	M	2	2	0
1710217	Q	5	4	1
1710218	O	5	4	1
1710219	N	2	2	0
1710220	L	4	2	2
1710221	F	1	1	0
1710222	X	0	0	0
1710223	D	5	6	–1
1710224	I	0	0	0
1710225	Z	2	2	0
1710226	E	2	0	2

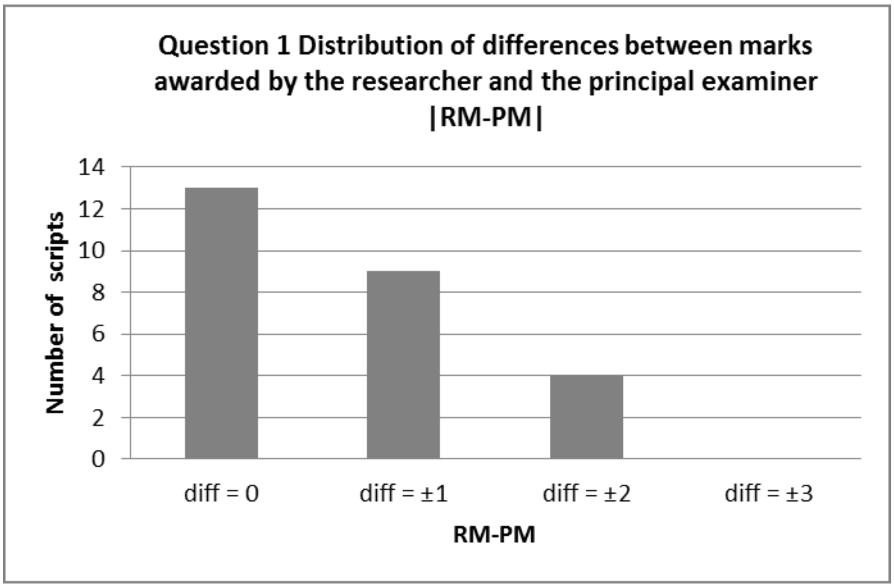


Figure 21 Distribution of differences between marks awarded by the researcher (RM) and the principal examiner (PM) |RM–PM|

The two sets of marks were compared using the same approach as in section 6.3. The level of agreement between the two sets of marks $P_{Agr1} = 0.84$ is similar to the agreement between the researcher and the original marker for this question ($P_{Agr1} = 0.88$). However the PE marks are on average a little lower than those of the researcher (RM-EM = 0.27), showing that the PE applied the theoretically-based mark scheme a little more harshly than the researcher. Kendall's tau coefficient $\tau = 0.817$ ($p< 0.001$) indicates that there is a strong correlation between the rank orders resulting from the marking of the researcher and the PE.

Normally an examiner would mark a set of training scripts and obtain feedback before proceeding to mark 'live scripts'. Given that there was no training before the marking took place, there is evidence from this trial that the mark scheme can be applied by other experienced markers without additional explanation.

6.5.2 Validation of the mark scheme for study question 2

Study question 2 was taken from the GCSE Physics paper (see appendix 1) and two principal examiners for physics agreed to trial the theoretically based mark scheme. The key data is given here with all the data provided in Appendix 4.

Table 13 shows the marks awarded by the two principal examiners (PM1 and PM2) and by the researcher (RM) and the difference between the researcher's marks and each of the principal examiner's marks. The distribution of the differences is show in Figure 22.

Table 13 Marks awarded by the researcher (RM) and the principal examiners (PM1 and PM2) for study question 2 during the validation process

Question 2 Nuclear power station		**mark awarded for question using mark scheme derived from theoretical framework**				**differences between marks awarded**		
candidate code	**script code**	**researcher mark (RM)**	**PE1 mark (PM1)**	**PE2 mark (PM2)**	**standardised mark (SM)**	**RM-PM1**	**RM-PM2**	**RM-SM**
1810201	B	2	4	3	3	–2	–1	1
1810202	D	2	1	2	1	1	0	1
1810203	F	2	3	2	2	–1	0	0
1810204	H	2	2	2	2	0	0	0
1810205	A	2	4	3	3	–2	–1	1
1810206	C	3	5	3	3	–2	0	0
1810207	J	2	1	2	1	1	0	1
1810208	E	4	4	4	4	0	0	0
1810209	T	4	4	2	3	0	2	1
1810210	G	3	2	1	2	1	2	1
1810211	S	3	1	1	3	2	2	0
1810212	P	3	3	3	3	0	0	0
1810213	I	3	3	2	2	0	1	1
1810214	K	5	5	5	5	0	0	0
1810215	R	5	5	4	5	0	1	0
1810216	M	5	4	3	4	1	2	1
1810217	Q	5	3	3	4	2	2	1
1810218	O	6	5	6	6	0	–1	0
1810219	N	4	4	3	3	0	1	1

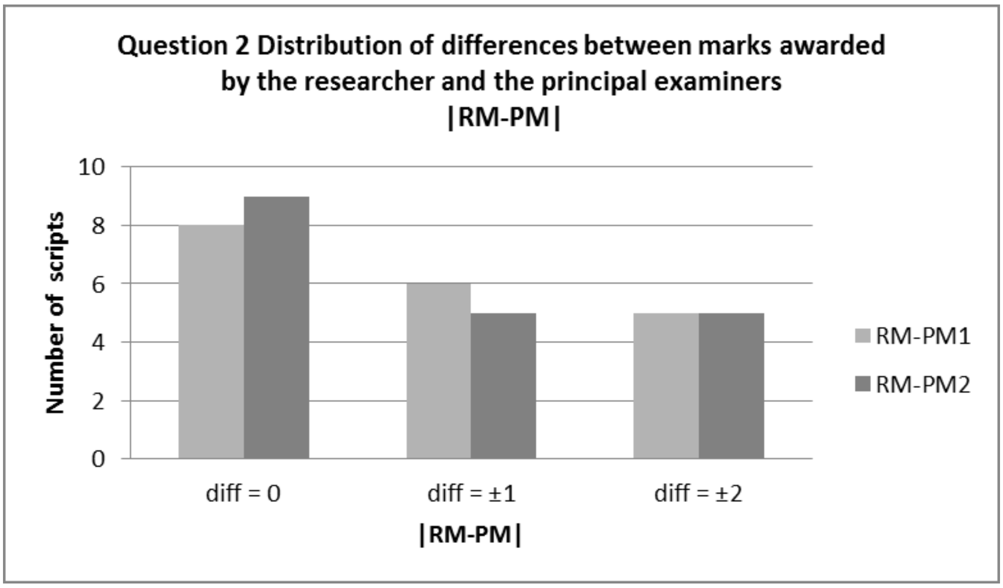


Figure 22 Distribution of differences between marks awarded by the researcher (RM) and the principal examiners (PM1 and PM2) |RM–PM1| and |RM–PM2|

The level of agreement between the each of the principal examiners scores and the researchers scores are both P_{Agr1} = 0.74, smaller values than the agreement measured for study question 1 (P_{Agr1} = 0.84). This lower level of agreement is reflected in the greater mean absolute differences, |RM–PM| = 0.79 and 0.84. The mean differences RM–PM = 0.05 and 0.53 show that both PEs applied the mark scheme more harshly than the researcher. Kendall's tau coefficient values $\tau = 0.506$ ($p< 0.01$) and $\tau = 0.489$ ($p<0.05$) shows that there is a weaker correlation between the rank orders given by the markers and the researcher compared with the PE who marked study question 1, perhaps reflecting a less well expressed mark scheme.

The researcher and two principal examiners met to discuss the mark scheme. The meeting took a similar form to a standardisation meeting. The three markers discussed their interpretation of the mark scheme and how they had applied it to each of the sample questions. After discussion a consensus was reached on the mark that should be awarded to each answer, these marks are identified as the 'standardised marks' (SM) in Table 13.

The data in Table 14 shows that for most candidates the marks awarded by the researcher and the two principal examiners were within 1 mark of the standardised mark (P_{Agr1} = 0.89, 0.95, and 1.00). The process of standardisation had produced a consensus, resulting in the rank order of candidates based on the standardised mark showing good correlations with each of the rank orders generated by the marks of the researcher and two PEs ($\tau = 0.70$, 0.75 and 0.76), which are close to the correlation between the rank orders produced by the researcher and the original examiner ($\tau = 0.75$).

Table 14 Effect of standardisation on marker agreement with standardised mark

Marker	P_{Agr1} (agreement with standardised mark)	Mean difference between marks awarded before standardisation and the agreed standardised mark	Mean absolute difference between marks awarded before standardisation and the agreed standardised mark
RM	1.00	0.26	0.58
PM1	0.89	0.21	0.63
PM2	0.95	0.47	–0.26

In the case of live examinations, after a standardisation meeting the scripts on which a consensus has been reached would be awarded those agreed marks and these questions would then provide models for markers when marking other scripts. Any points arising from the discussion would be used to annotate the mark scheme to ensure that everyone applied it in the same way.

In this part of the study the fact that the three markers were able to come to a consensus that was not far from the original marks awarded by each marker suggests that a clarification of the mark scheme during discussion was needed to provide the common understanding.

6.6 Summary

The work reported in this chapter has shown that it is possible to develop a theoretically-based framework that can be used to construct mark schemes that yield comparable marks with the original marks on a range of questions. It has also shown that experienced examiners can apply the theoretically-based mark schemes to two of the questions, yielding marks that are close to those awarded by the researcher.

When there was an opportunity to discuss the mark scheme with the examiners there was even closer agreement on how the mark scheme should be applied, reflecting the process that takes place during standardisation.

The next chapter describes result of sharing the ideas of the study with a broader range of assessors and canvassing their views of the usefulness of the approach.

Chapter 7 Findings from assessors' responses to the study

7.1 Introduction

This chapter reports on the qualitative data collected in this study to answer research question 3: *Would examiners find such a framework [i.e. a framework based on Erduran et al.'s interpretation of Toulmin's argumentation pattern] useful in establishing consistency of demand across different papers within the suite and year on year?*

Section 7.2 summarises the outcomes of the discussion with the senior examiners who had previously marked study question 2 (section 6.3). The results of that interview informed the development of the questionnaire which was sent to all the assessors for OCR GCSE Science A. Section 7.3 gives the rationale for the questions used in the questionnaire and reports on the answers of those who responded.

7.2 Discussion with principal examiners

The 'standardisation' meeting with the two principal examiners who had marked study question 2 as part of the validation process is described in section 6.5.2. On the occasion of that meeting the researcher discussed the principles behind the study with the two principal examiners (identified as PE1 and PE2 in this chapter). This conversation took place in March 2013 at which time there had been three examination series for GCSE science papers which included levels-based mark schemes, so the two examiners were becoming familiar with the process of writing levels-based mark schemes and had also used the schemes to mark live examinations. The discussion was recorded and this report is based on notes made during the conversation, backed up by the recording.

The PEs were asked about how much tolerance they felt was appropriate when comparing the marks awarded by different examiners, for levels-based marks schemes. They agreed that there needed to be some tolerance when the marking involves a degree of judgement by the examiner. There was agreement that that a tolerance of ±1 on each 6-mark question would be acceptable, which is compatible with the approach used in this study to measure reliability of the mark scheme.

The examiners were asked for their thoughts about using a theoretical framework based on grade descriptors as the basis for a mark scheme. They agreed that in principle it could be an extremely useful starting point for developing mark schemes. PE2 pointed out, however, that some questions on an examination paper may prove to be more challenging than anticipated and consequently it may be necessary to ease the marking of a challenging question to try to ensure that there is a good differentiation between candidates; if a question is too demanding then the upper part of the mark range is not used.

Another concern expressed by PE2 was that the framework would not work for all the questions that require extended answers: "I think this is a good idea, though I am not sure how well it fits for the full variety of questions." PE1 agreed, "it is helpful, yes, but there may be questions where it may drive us to apply a framework which does not necessarily apply."

The current specification does not use the language of argument expressed in the framework, in particular PE2 pointed out that "we don't ask for rebuttals in questions that ask for the arguments". PE1 responded "you might decide that if we were to use this framework we might cue them better, if you are asking for different options, you can ask for arguments against the alternatives".

The researcher pointed out that in the framework (Table 5) it is only at level A3 that a rebuttal is absolutely required, below that level an alternative to a requirement for a rebuttal is for candidates to write about the quality of the evidence. PE2 said "that is really assessed in the coursework, but it could also be assessed in exams".

Overall the response of the two examiners was that the framework could prove to be a useful tool for principal examiners, though they would want to retain the flexibility to adapt to the circumstances.

7.3 Questionnaires

7.3.1 Development of the questionnaire

A questionnaire was devised to explore the views of those involved in the development of examination questions and mark schemes on the framework developed in this study. The questionnaire, accompanying letter and description of the study are included in Appendix 5. The questionnaire was sent out in May 2013, by which time those involved in setting papers had written six sets of papers and marked papers from three live sessions.

This section explains the rationale for each of the questions asked in the questionnaire. (Although the mark schemes developed in this study are described as levels-based mark schemes in this report, amongst the GCSE science community at OCR they are known as 'level of response' (LOR) mark schemes, and that is how they are referred to in this questionnaire.)

1 Do you think that questions with LOR mark schemes allow examiners to assess skills that are not rewarded using 'conventional' extended answer mark schemes?

Rationale This question was intended to find out how the assessors perceived levels-based mark schemes at a time when the schemes had been in use for three live sessions. Any future developments of levels-based mark schemes are more likely to be successful if the assessors have a positive attitude to them.

2 Do you think that the challenges of writing an LOR mark scheme are different from those met when devising a mark scheme for other questions that require an extended written answer?

Rationale This question was intended to find out how the assessors perceived the challenges of writing such mark schemes after they had the experience of writing them for six sets of papers. If there are particular challenges, any development to the schemes should try to address those challenges.

3 Do you think that starting from a common framework based on the grade descriptors would be helpful in devising LOR mark schemes?

Rationale This question was intended to find out whether a common framework based on grade descriptors would help in writing mark schemes and whether assessors would mention, without being prompted, that it could increase comparability of demand between papers within the suite.

4 Do you think that the framework devised in this research specifically for questions that ask for explanations or arguments could be useful?

Rationale This question was intended to find out whether assessors can see benefits in frameworks for particular questions. If they can see some benefits of the change, it would be easier to implement it.

7.3.2 Target population for the questionnaire

It was intended that all those involved in the question and mark scheme development process (described in this study as the assessors) would be invited to answer the questionnaire in May 2013. As OCR could not disclose the contact details for all those people, a member of the qualifications team at OCR sent the questionnaire, covering letter, and description of the study to 25 people in the target population, in total 12 people responded (although one response was an apology for not answering all the questions owing to illness). Several of the respondents have more than one role and between them they cover all the different roles involved in the setting and marking process; one was a chair of examiners, seven were principal examiners (of whom four were also chief examiners) who set papers, eight were principal examiners who lead the marking of papers, three were revisers, and one was a scrutineer. Some of the respondents also have experience of these roles for other OCR specifications or with other awarding bodies. For more about these roles, see section 2.3.

Unfortunately the timing of the questionnaire survey coincided with the beginning of the summer examination session, a very busy time for examiners, which may partly explain the disappointing response rate. Many commentators suggest that a response rate of 60% to a survey is acceptable (Robson, 2011). In this study the response rate was only 48% of the whole population and, although all the roles in the setting and marking process are covered, the responses should not be considered to be representative of the views of all assessors.

7.3.3 Responses to the questionnaire

With the small number of responses to be considered it would not be sensible to attempt any kind of statistical analysis, and so the report in this section is simply qualitative. The section summarises the responses to each of the questions asked together with some examples of typical responses. To preserve anonymity the various roles of the respondents are not given, the individual respondents are identified as Assessor1, Assessor2, etc..

1 Do you think that questions with LOR mark schemes allow examiners to assess skills that are not rewarded using 'conventional' extended answer mark schemes?

All those responding gave a positive response, making the point that it is important that candidates can express their science knowledge clearly. Assessor1 said that a points-based mark scheme does not have scope to take account of the way an answer is expressed:

> the organisation of a candidate's answer is important and their ability to link ideas logically. With conventional mark schemes, answers that had the correct points may gain full credit, even if the ideas were not properly logically linked.

Assessor2 also liked the fact that levels-based mark schemes can reward a candidates' ability to communicate science:

> LOR mark schemes allow for effective assessment of the quality (as opposed to the quantity) of the answer, in that it can avoid a sliding scale derived from a 'points – based' approach and give scope for crediting the answer in terms of the complexity of the response.

2 Do you think that the challenges of writing an LOR mark scheme are different from those met when devising a mark scheme for other questions that require an extended written answer?

Almost all those who replied answered 'yes' to this question, with two people responding 'yes and no'. Assessor3 summarised the issues by saying:

> The problem is to produce a mark scheme that retains its accuracy and reliability while at the same time assessing new skill areas.

Assessor1 focused on the need for writing a mark scheme that can be used reliably:

> Making the mark scheme brief enough to be useable, given that we are trying to build in as much range as possible. Making each level clearly discrete so that it is clear how L2 is diff to L1 and L3 is diff to L2 so that marking is consistent.

The assessors who responded 'yes and no' made the point that there is still a need to identify the science content that is expected, but that there is the additional challenge of describing performance. Assessor4 said:

> There is a heavy degree of overlap [in the challenges of writing the different mark schemes], but there are also differences ………. the need to balance the competing demands of argument and facts for questions which ask candidates to explain a situation with specific reference to given piece of information / theory.

3 Do you think that starting from a common framework based on the grade descriptors would be helpful in devising LOR mark schemes?

Most assessors responded 'yes' to this question, though some had some caveats. Two responded 'no'. Amongst those who responded positively, a common theme was the recognition that it would help to bring consistency between subject teams:. For Assessor5 this is currently a concern:

> I worry about divergence between Biology, Chemistry and Physics now that we operate in separate little bubbles.

Assessor6 responded very positively to the framework and recognised the potential for its use to lead to greater consistency, and also suggested that it may improve the questions too:

> I like the idea – because it is principled – and could lead to greater consistency. Also it might lead to better questions more likely to elicit the kinds of response that we should be looking for.

Assessor7 suggested that such a framework might also help teachers to understand better what examiners are looking for. He/she commented that:

> having a standardised LoR base would be helpful for the examiners, and also for teachers. …………… However, I think that there would need to be some support materials available for teachers (and examiners).

Assessor3 raised a concern that had been mentioned in discussion with the PEs (section 7.2); using the framework should not limit the examiners, because

> the (Principal) Examiner is often forced to tighten or relax the mark scheme to achieve a distribution of marks that aids awarding and give a good distribution of marks. This flexibility would be removed.

However Assessor8 recognised this issue and suggested that the framework could still be used, but using a different range of descriptors, and acknowledging that a different target grade is appropriate:

> there is a concern that the mark schemes have to be altered in the light of candidate responses. Not sure if this matters as they can still be based on grade descriptors – it just means the original marking grid of [for] the paper was wrong.

Assessor9 did not think that linking the mark scheme to the grade descriptors would be helpful:

> I don't find grade descriptors very helpful as a starting point. They are more useful when assigning grades after the question and mark schemes have been used.

The use of the descriptors at the awarding stage mentioned by Assessor9 above was considered by Assessor2 to be a positive feature of the framework:

> it helps to match the mark scheme to the perceived range of levels of demand in the question. From an awarding point of view, this helps in matching key boundary performance against the grade descriptions.

4 Do you think that the framework devised in this research specifically for questions that ask for explanations or arguments could be useful?

All respondents gave positive answers to this question, though again there were some caveats. Assessor8 suggested that it would help to establish consistency between examiners:

> It gives 3 logical levels. Markers should be able to apply this easily and it would give consistency in the marking.

Assessor10 made that the point that starting from the same framework would help mark schemes become more familiar and so lead to more consistent marking.

> It defines a different level of response for each level, so becomes clearer which level an answer is in, and then choosing whether it is the higher or lower mark is also easier. ……….to get everyone to mark consistently a framework that everyone understands is the key – not a different one for each exam question.

Assessor2 said that the framework would

> help to establish the validity of the question and its MS in testing the required assessment objectives.

Assessor4 suggested that the approach might be considered for levels-based mark schemes that address other types of extended writing:

> Yes, very helpful for this style of LoR question. It may be worth surveying the range of LoR question styles which are currently being produced and devising strategies for each different approach.

Assessor4 went on to express the concern that the framework may become a requirement that is imposed:

> I have one major fear, and that is that a well-meaning officer of the Board will demand that all LoR mark schemes fit your framework[s], and so create a constraint rather than a helpful tool.

Assessor1 reiterated the concerns made earlier by Assessor3 that Principal Examiners need freedom to amend mark schemes at the standardisation meeting:

> …. it is important that we are 'free' at SSU [standardisation meeting] to look with an open mind at the range of candidate responses and design a MS to suit. The MS is often very different from the outline we envisaged through AMEC [QPEC]. We are already constrained by a great deal of (Ofqual and self-imposed!) rules and limitations on how we can approach 6 markers so I think it's important that we don't self-impose more.

A number of the assessors made the point that the original introduction of levels-based mark schemes for the GCSE Science examinations from 2012 onwards was made without much training for examining team – or for teachers. Assessor6 made it clear that if a new framework were to be introduced then there should be training for both examiners and teachers:

> It looks promising but needs to be tested in more detail. It is not only examiners that would need training – but also teachers. Students have to be prepared to present high quality arguments in their answers. To do so they need to know the features of good arguments – including the importance of 'rebuttals' which seem to be given particular prominence in your examples.

7.4 Reflections on the assessors' responses

Overall the assessors were positive about levels-based mark schemes and most could see the usefulness of a framework that provided a starting point for developing mark schemes for specific questions.

Research question 3 is intended to explore potential of the framework to improve consistency of demand between papers and across time. The questionnaire did not ask about this directly, as it would have been difficult to do so without it becoming a leading question. A number of responses did recognise the potential of a framework to provide consistency between subjects and sessions; the benefit of knowing the marking demand of questions at the grade award stage was also mentioned.

Several of the assessors made the point that they do not want to be constrained by a framework which would restrict their ability to adapt the mark scheme at the standardisation meeting, based on the sample of candidates' work. There does not need, however, to be a conflict between maintaining consistency of demand and allowing flexibility at standardisation, provided there is flexibility to move up or down the level descriptors 'ladder' and that information about the change in demand is available at the grade award.

Almost all the 25 assessors to whom the questionnaire was sent are known to the researcher to some degree. It may be that those who did not respond felt less positively about the study and found it difficult to say that to a researcher they know, if only slightly. For this reason, care must be taken not to overestimate the significance of the positive responses described here; however the respondents who replied all seemed to think the approach was useful and could improve some aspects of the examination process.

Chapter 8 Conclusions

8.1 Overview

The purpose of this study was to find out if it was possible to develop a generic framework that could be used to write mark schemes for questions in GCSE Science that required an extended response that incorporated an argument or scientific explanation. If this framework were to be useful it would need to be applicable to a range of questions of this type, the mark schemes would need to capable of being used successfully by other examiners, and the principles behind the development would have to be understood and accepted by the senior assessors who might be asked to used such a framework.

A generic, theoretically-based framework was developed to describe and evaluate student's answers to such questions; this was used to write mark schemes for questions that had been used in GCSE Science examinations. These mark schemes were used to mark students' answers and the marks were compared with those awarded using the original mark schemes. The utility of the mark schemes was confirmed by checking that other examiners could apply the mark scheme, yielding similar results. Those involved in the examining process for the GCSE Science specification used in the study were invited to respond to the ideas developed.

This chapter reviews the research questions that guided the study and considers what conclusions can be drawn from the findings of the study. It goes on to evaluate the research methods and strategy and suggest how they might be improved if further work were to be carried out. Section 8.5 considers the implications of the study and how its findings could inform future practice in setting and marking examination questions. Finally there are suggestions for how the work might be extended by widening its scope.

8.2 Answering the research questions

Q1 Can a theoretically-based model of argumentation be used as the basis for developing a framework to evaluate extended answers to questions in GCSE Science examinations that require an argument or explanation?

The theoretical basis of this study was the model of argumentation of Toulmin (2003). Erduran et al. (2004) used Toulmin's model to write an analytical framework for evaluate argumentation in the science classroom (Table 3). Section 5.2 of this report described how the grade descriptors for GCSE Science (Ofqual, 2009) were mapped to the descriptors used by Erduran et al. This mapping was used as the basis for a generic framework of level descriptors of the type used in levels-based mark schemes for GCSE Science questions that ask for explanations or arguments. This generic framework was used to write mark schemes for five specific questions from the OCR GCSE Science A examinations.

Q2 How do the marks awarded when using such a framework compare with those awarded using the conventional mark scheme from the awarding body?

As shown in Table 11 (section 6.4), the overall the agreement between marks awarded by following the two schemes was good (mean $P_{Agr1} = 0.89$). The mark schemes generated by the study yielded marks on average -0.3 marks (out of 6) lower than the mark schemes that were used by the examiners, which might suggest that the study mark schemes were setting a more demanding standard than the original schemes, but further investigation would be needed to know why the original marks were awarded. There were strong correlations in the rank orders between the marks awarded using the conventional mark schemes and the research-based schemes (range $\tau = 0.61$ to 0.83, $n = 19$ to 26), which implies that the mark schemes produced using the theoretically-based framework would yield similar outcomes for candidates.

Q3 Would examiners find such a framework useful in establishing consistency of demand across different papers within the suite and year on year?

Although the research question refers to examiners, the questionnaire used in the study solicited the reactions of a broader cohort of assessors involved in the examination process, including chairs of examiners and revisers. The response of the assessors was generally positive, with a number spontaneously identifying the potential of the framework to provide some consistency between subjects and between examination sessions. One assessor returned the questionnaire with the comment that “this is a very pertinent piece of research for Twenty First Century Science” (Assessor2).

Some of the examiners who replied positively had some reservations about a framework being imposed because they would want to retain the flexibility to change the demand of a mark scheme with the aim of increasing the spread of marks awarded across the paper. These concerns can be answered by pointing out that basing the mark scheme on a framework which makes explicit the levels of demand could make a change of this kind more transparent. Examiners wishing to change the demand of the mark scheme could do this by using a different part of the framework.

When used by other senior examiners the schemes yielded similar marks, (P_{Agr1} = 0.89 and 0.95), which shows that the mark schemes can be used reliably by other examiners. The idea of a common framework for the development of mark schemes was seen as a useful idea by those assessors who responded to the questionnaire.

Overall it can be concluded that, within the limitations of the study (see section 8.4), mark schemes for questions that require candidates to give an explanation or make an argument can be written using the framework developed in this study. Evidence for this is that the theoretically-based mark schemes yielded similar outcomes to the original mark schemes used to mark the questions (mean difference between marks awarded originally and using the theoretically based scheme was -0.3 for these 6 mark questions (range 0.00 to -0.65); the rank orders were similar – Kendall’s tau correlations ranged from 0.61 to 0.83). These moderate correlations suggest that whilst there was not an exact match in outcomes for individual candidates the two mark schemes were measuring similar things. If that is the case, it might be argued that there is no need to change from mark schemes written by examiners using their professional experience and intuition about what makes a suitable answer at a particular grade, to mark schemes based on a theoretically-based framework. On the other hand, using a mark scheme based on a common framework, that is grounded in theory and takes account of the grade descriptors, would support the awarding of grades (see section 2.6), by providing places in the paper where the examiners can show that candidates’ performance has been marked with those grade descriptors in mind. Using the framework across a suite of specifications would provide a way of looking for comparable outcomes across different papers and different sessions.

8.3 Evaluation of the methodology and strategy

8.3.1 Using the theoretically-based framework to write and test mark schemes

The study used five questions that asked for explanations or arguments from examination papers for GCSE physics and GCSE chemistry, but did not use any questions from examinations for GCSE biology. This was because the Item Level Data for the possible biology questions showed that those questions had not discriminated well between candidates.

Now that the framework has been shown to work for physics and chemistry questions, it would be important to check that the framework can also be used to develop and test mark schemes for biology questions before recommending its wider use. There have been three further examination sessions since the study began, which may yield suitable questions for this check.

The size of the samples of scripts used was limited by the availability of material from OCR. A larger number of scripts for each study question would provide the opportunity to improve the validation process described in section 6.5 by including an extra stage. Following the 'standardisation discussion' between the researcher and two principal examiners a further batch of scripts could be marked by all three examiners. If the discussion really had brought better understanding of the scheme it would be expected that there would be an even closer agreement about the marks awarded to each answer.

8.3.2 The validation process

The process of checking the validity of the mark schemes (section 6.5) included asking senior examiners to mark the questions without any training or exemplar scripts. This might seem to reduce the validity of the check, as in normal circumstances markers receive training before using a mark scheme. However Baird et al. (2004), have found that experienced examiners were able to mark reliably without exemplar scripts or training (see section 3.2.3).

8.3.3 The questionnaire

The response rate for the questionnaire was disappointing. This may be partly attributable to the timing of its distribution, at the beginning of the examination session when the Principal Examiners were busy preparing to lead marking teams. An additional problem was that the initial letter did not specify a return date; a deadline might have encouraged busy people to respond quickly. A subsequent reminder email did include a suggested return date and yielded a few more returns.

If the questionnaire had been sent at a less busy time of year it might have yielded a higher response rate. An alternative approach that might be expected to produce a higher response rate would be to take the opportunity of a face to face meeting to explain the research and present the questionnaire, or alternatively to administer the questionnaire as a structured telephone interview.

8.4 Limitations of the study

A limitation of this study is that although the mark schemes developed from the theoretical framework were used successfully by other examiners, the framework itself was not used by other examiners to write mark schemes. Any further development of this work should begin by checking that others can apply the framework to write mark schemes that can then be used by others.

The framework used in the study was developed for specific questions types within one specification suite. It cannot be assumed that the same framework could be used by others, though the principle of using a model of student learning alongside grade descriptors perhaps has the potential to be used more widely. An alternative approach to writing level descriptors using empirical evidence from candidates' work is described by Greatorex (2003), this method would not have been possible for the examination questions used in this study as the style of question used in the study was being examined for the first time in the January 2012 session .

It might be argued that using the Toulmin model of argumentation as a starting point for the study is flawed because the formal language used by Toulmin to describe arguments (warrant, qualifier, backing, and rebuttal) is not part of the language normally used in describing quality of arguments and explanations in GCSE Science. However candidates are expected to use argument and explanations in their answers and the mark schemes were adapted to accommodate this, as described in section 5.2. The advantages of testing the framework on candidates' work in GCSE examinations (see section 4.4.2.) outweighed the alternative of preparing students to answer questions that tested their ability to construct an argument that followed the Toulmin model of argumentation.

8.5 Implications of the findings

8.5.1 Assessment of argumentation

Many educators believe that the role of argumentation in the development of scientific ideas should be taught in science lessons and this study has shown that it is possible to write a levels-based mark scheme that rewards answers that use the elements of argument. However whilst current GCSE Science specifications make mention of the process of argument they do not make explicit exactly what is required to make a good argument. The developers of the National Curriculum programme of study for science, the subject criteria for science, and science specifications, should be encouraged to include the role argument in the practice of scientific more explicitly in their documents.

These ideas may be new to some teachers, and questions that asked students to present an argument supported by evidence would help operationalise this aspect of the specification for teachers (Millar, 2013). Mark schemes based on a common theoretically-based framework would make clear what is expected in answers. Making the frameworks and information about the theoretical background to the frameworks available to teachers should help those teachers to appreciate the underpinning ideas on which mark schemes are based, and consequently to develop suitable teaching approaches. Assessing the practice of argumentation is, perhaps, the surest way of ensuring that the practice is taught in schools, teachers will teach what is tested (Baird et al., 2013).

8.5.2 Using theoretically-based frameworks

The success of the framework in devising useable mark schemes and the generally positive feedback from assessors suggests that OCR, and other awarding bodies, might explore the idea of asking GCSE Science examiners to use the grade descriptors that are part of the subject criteria (Ofqual, 2009) when writing levels-based mark schemes. This section considers the implications of this suggestion for both assessors and teachers.

Currently the expected answers in mark schemes are based on the experience of the assessors and the requirements of the specification. If it were agreed that the principles behind the development of this framework should be applied more broadly to the writing of mark schemes, there would need to be a major change in practice for assessors. This was recognised in some of the feedback to the questionnaire reported in section 7.3.3.

It can be difficult to bring about a change in practice by professionals who have been working in a particular way for many years (all those who responded to the questionnaire (n = 12) had been assessors for GCSE Science for more than 6 years, seven of them for more than 10 years). To make such a change successful, those who have to alter their way of working must see how it will benefit their practice; they must understand how the change can be brought about and believe it to be manageable; and they should believe that the change would bring worthwhile improvements (Fullan, 2007). In writing about why teachers do (or do not) embrace change Doyle and Ponder (1977) refer the ‘practicality ethic’ – which they suggest has three dimensions: instrumentality, congruence and costs.

Each of these would need to be considered by awarding bodies if they were to implement the proposed change in practice.

> **Instrumentality** – the examiners need to see how they would implement any proposals. This would need careful management – if they feel that the proposals are imposed from above examiners may resent the imposition (a concern voiced by Assessor4 in section 7.3.3). It is recommended that at least some of the examiners should be involved in the development of the

framework of descriptors. It would be necessary to provide training for all the examiners so that they come to a common understanding of the purpose and use of such frameworks to develop mark schemes.

Some of the assessors who responded to the questionnaire recognised the need for such development and training in their responses to question 4, Assessor1 commented that "this is a useful approach…it would be a great approach to use in training and for reflection"

Congruence – the examiners need to see how a new way of writing and using mark schemes would fit with their current practice. For example, those examiners who raised concerns about the need retain the flexibility to change mark schemes at the standardisation stage should be shown how the demand of the scheme could be changed whilst still maintain its integrity within the framework. At least some of the assessors recognised that flexibility (see the comment by Assessor8 in response to Question 3 in section 7.3.3), so it should be possible to show others that it is possible.

Cost – examiners would need to understand how the benefits of using a theoretically-based framework would be worth the cost of changing their practice and perhaps the cost of feeling they have less control of their own work. This could be problematic because many of the benefits identified by the assessors and reported in section 7.3.3 do not come to the individual examiner and his/her team. The potential of a theoretically-informed mark scheme to support the awarding process was identified by one of the assessors, and another reflected on the opportunity the framework would provide to improve consistency in standards between the papers set for the three sciences. These benefits might be seen to improve the validity of the assessment, but bring no direct benefit to the individual examiners. For these examiners the most obvious benefit of a suitable framework might be show how it would give them a starting point for writing specific mark schemes, which would in time have a familiar structure for examiners.

The current specifications for GCSE Science will be examined for the last time in June 2017. Papers have been written for the 2016 session, so there is only more set of papers to be written. This would not be a sensible moment to introduce a change such as the one explored in this study. The next section considers how the work from this study might inform future developments in GSCE Science.

8.6 Further work

This study has shown that it is possible to develop a mark scheme based on a theoretical model of the structure of argumentation in science. Whilst such mark schemes would not significantly alter the rank order of the candidates, they would provide a more transparent basis for allocating marks, evidence to support grade awards, and an opportunity for increasing consistency of standards across subjects and examination sessions. A sample of experienced assessors were positive about the approach, whilst at the same time identifying some concerns that would need to be addressed if it were to be taken further.

The principle of aligning the mark schemes more closely with the grade descriptors was identified as a positive aspect of the framework developed in this study. The regulatory system for GCSEs is currently under discussion with the reformed GCSEs in Science to be examined for the first time in June 2018 (Ofqual, 2013b). This reform process includes the introduction of a new grading system (from 1-9, rather than the current G-A*), this will require new grade descriptors which will be needed to set the standards of the new examinations.

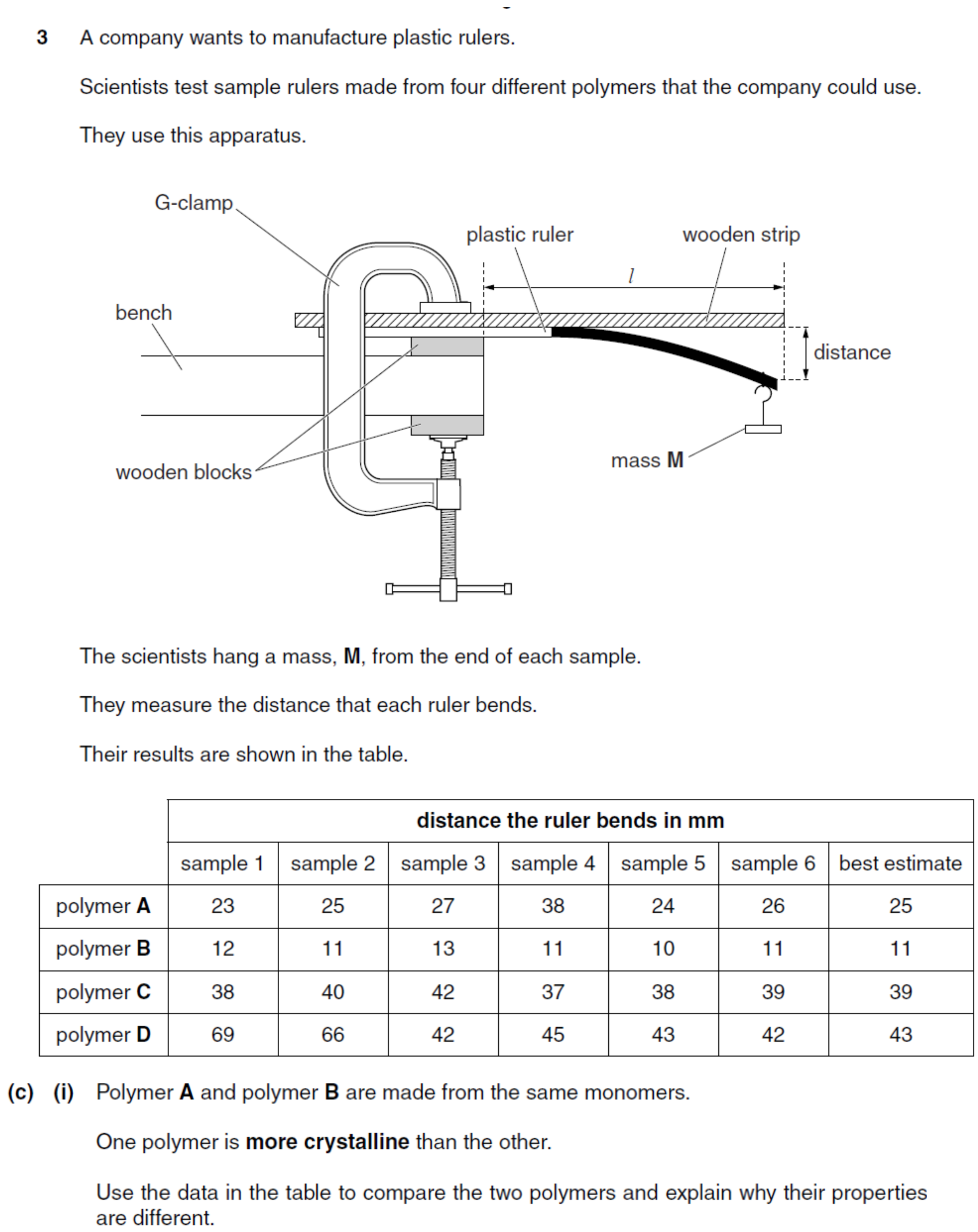
The work of this study could be extended to develop frameworks suitable for each of the main questions types used in GCSE Science extended response questions and linked to the new grade descriptors. Forging close links between the grade descriptors and the questions and mark schemes would not only help to ensure that examiners engage with the new grade descriptors at an early stage, but would also help to demonstrate the relationship between the assessment and the specification. Such a framework would also be very useful for teachers. The descriptors would help operationalise the specification, particularly if the framework were accompanied by a series of sample questions and mark schemes that show the relationship between the assessment objectives, the grade descriptors and the science content.

The benefits of taking a more principled approach to writing questions and mark schemes, and that have been identified in this study, make it worth exploring ways in which the work could be taken further in the way outlined above.

Appendix 1 Questions and mark schemes used in the study

Question Identifier	**Paper and question number**	**Topic of question**	**Level of demand**
Q1	GCSE Chemistry A171/02 Q3ci (OCR, 2012i)	Properties of polymers	A
Q2	GCSE Physics A181/02 Q6 (OCR, 2012r)	Siting a nuclear power station	B
Q3	GCSE Chemistry A171/01 Q2b (OCR, 2012h) and A171/02 Q1b (OCR, 2012i)	Particulates and asthma	C
Q4	GCSE Physics Physics A181/01 Q6(OCR, 2012q) and A181/02 Q4 (OCR, 2012r)	Risks of sunbathing	C
Q5	GCSE Physics Physics A181/01 Q9 (OCR, 2012q)	Siting of a power station	D
Q6	GCSE Biology Biology A161/01 Q5 (OCR, 2012d)	Environmental indicators	D

Study Question 1 Polymers A171/2 Q3(c)(i)



3 A company wants to manufacture plastic rulers.

Scientists test sample rulers made from four different polymers that the company could use.

They use this apparatus.

G-clamp
plastic ruler
wooden strip
l
bench
distance
wooden blocks
mass **M**

The scientists hang a mass, **M**, from the end of each sample.

They measure the distance that each ruler bends.

Their results are shown in the table.

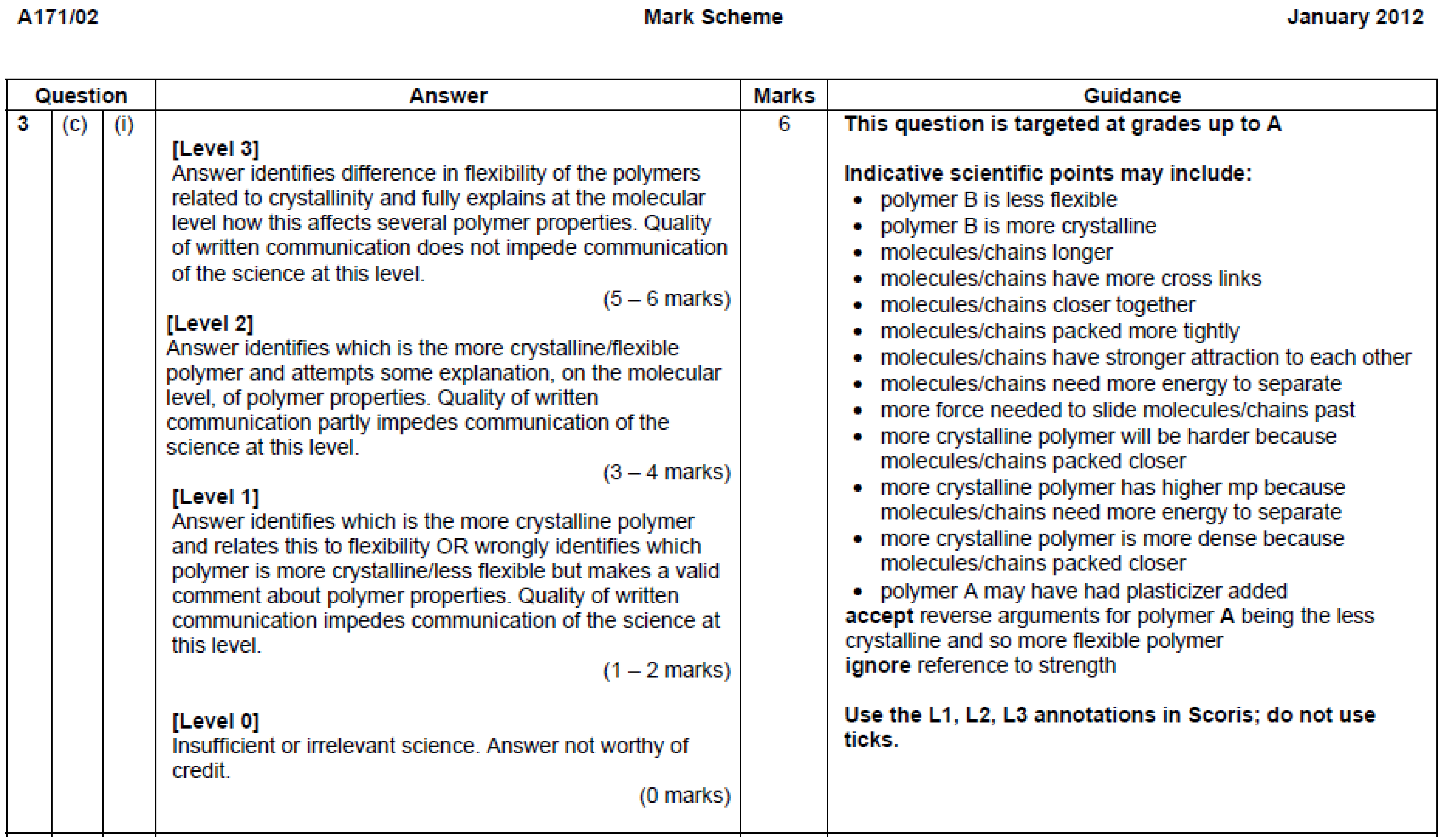
	distance the ruler bends in mm						
	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	best estimate
polymer **A**	23	25	27	38	24	26	25
polymer **B**	12	11	13	11	10	11	11
polymer **C**	38	40	42	37	38	39	39
polymer **D**	69	66	42	45	43	42	43

(c) (i) Polymer **A** and polymer **B** are made from the same monomers.

One polymer is **more crystalline** than the other.

Use the data in the table to compare the two polymers and explain why their properties are different.

Figure 23 Study question 1 A 171/2 Q3(c)(i) (OCR, 2012i)k



A171/02 Mark Scheme January 2012

Question			Answer	Marks	Guidance
3	(c)	(i)	**[Level 3]** Answer identifies difference in flexibility of the polymers related to crystallinity and fully explains at the molecular level how this affects several polymer properties. Quality of written communication does not impede communication of the science at this level. (5 – 6 marks) **[Level 2]** Answer identifies which is the more crystalline/flexible polymer and attempts some explanation, on the molecular level, of polymer properties. Quality of written communication partly impedes communication of the science at this level. (3 – 4 marks) **[Level 1]** Answer identifies which is the more crystalline polymer and relates this to flexibility OR wrongly identifies which polymer is more crystalline/less flexible but makes a valid comment about polymer properties. Quality of written communication impedes communication of the science at this level. (1 – 2 marks) **[Level 0]** Insufficient or irrelevant science. Answer not worthy of credit. (0 marks)	6	**This question is targeted at grades up to A** **Indicative scientific points may include:** • polymer B is less flexible • polymer B is more crystalline • molecules/chains longer • molecules/chains have more cross links • molecules/chains closer together • molecules/chains packed more tightly • molecules/chains have stronger attraction to each other • molecules/chains need more energy to separate • more force needed to slide molecules/chains past • more crystalline polymer will be harder because molecules/chains packed closer • more crystalline polymer has higher mp because molecules/chains need more energy to separate • more crystalline polymer is more dense because molecules/chains packed closer • polymer A may have had plasticizer added **accept** reverse arguments for polymer **A** being the less crystalline and so more flexible polymer **ignore** reference to strength **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**

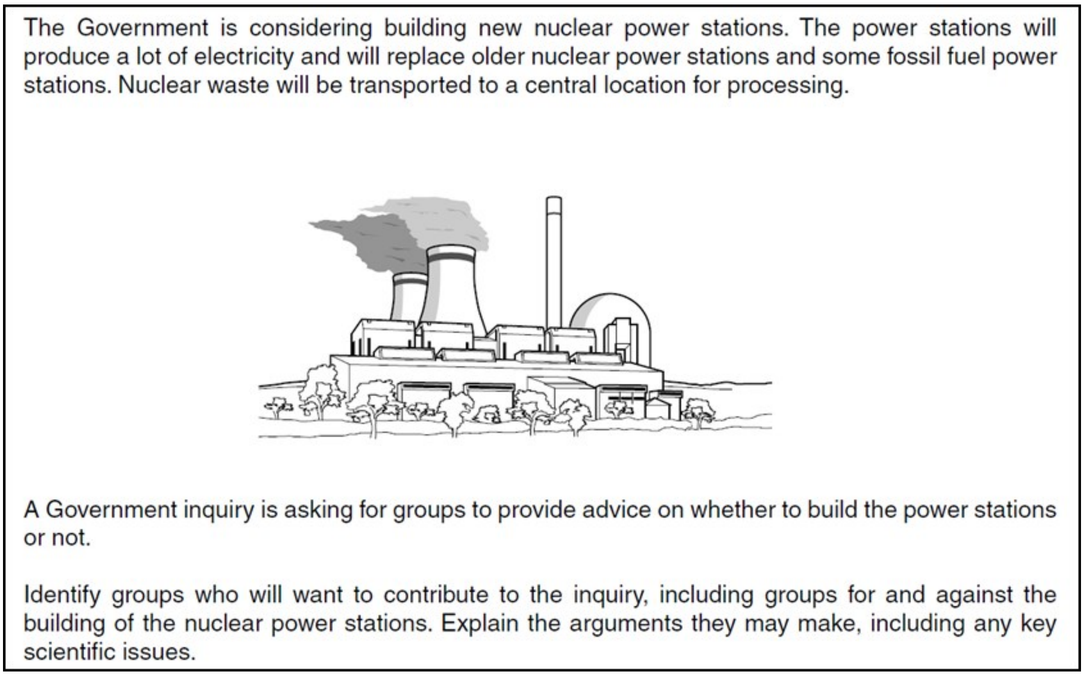
Figure 24 Study question 1 OCR Mark Scheme A171/2 Q3(c)(i) January 2012 (OCR, 2012i)

Study mark scheme Question 1 Polymers– A171/2 Q3(c)(i)

Framework descriptor	**Mark scheme descriptors**	**Guidance**
A3 The argument or explanation of a claim is supported by evidence (data) with clear scientific reasoning (warrant and backing). The argument takes account of the limitations of the evidence or provides a rebuttal to possible counterarguments. No serious errors of science.	**Level 3** Makes correct **claim** Gives **evidence** **Reasoning** links molecular structure to stiffness and to other properties No serious science errors. Quality of written communication does not impede communication of the science at this level.	This question is targeted at grades up to A. Throughout the candidate could make a reverse argument in terms of why A is less stiff. (If candidate does not identify the correct polymer they may score up to level 2 for correctly linking structure and properties.) **Claim** • Polymer B is more crystalline **Evidence** • Sample B deflects less under load • So is stiffer **Reasoning** • In crystalline polymers o polymer chains are ordered with cross-linking o polymer chains are more closely packed o polymer chains have stronger attraction to each other • so chains more difficult / need more energy to separate • so chains more difficult / need more energy to slide • (so more stiff) Other differences in properties due to crystallinity: More crystalline will be: • denser • harder • higher melting point Possible **Rebuttal A** could be more flexible because it includes plasticisers, rather than less crystalline. However the question does not lead naturally to rebuttals as the data is clear cut.
A2 The argument or explanation of a claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence / argument.	**Level 2** Makes **claim** Gives **evidence** **Reasoning** describes or explains some aspect of the link between molecular structure and stiffness, without the complete argument. Quality of written communication partly impedes communication of the science at this level.	
A1 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.	**Level 1** Makes **claim** Gives **evidence** Makes some further comment about structure/behaviour. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Figure 25 Study question 1 Mark scheme derived from theoretical framework

Study Question 2 Nuclear power station A181/2 Q6

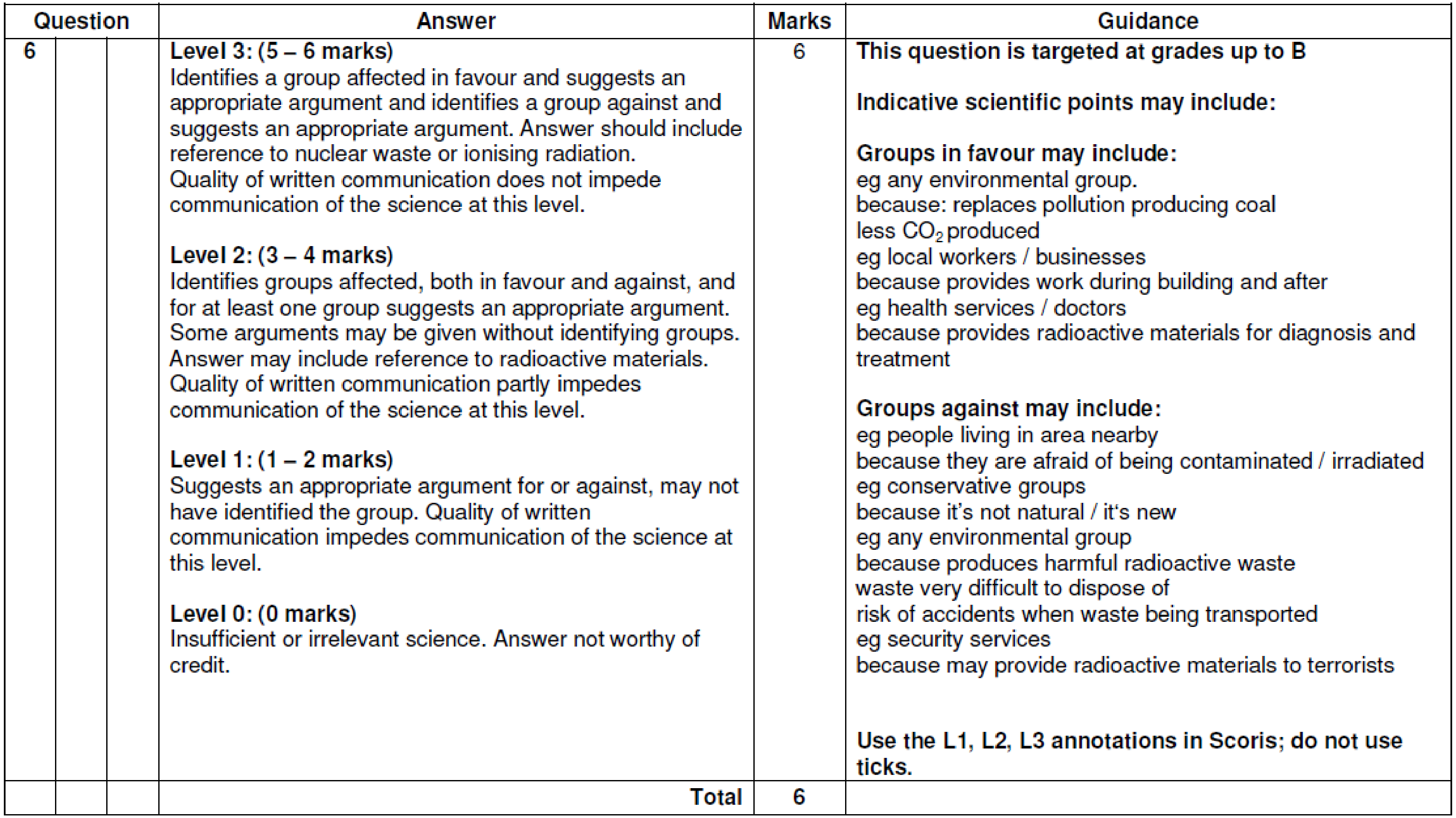


The Government is considering building new nuclear power stations. The power stations will produce a lot of electricity and will replace older nuclear power stations and some fossil fuel power stations. Nuclear waste will be transported to a central location for processing.

A Government inquiry is asking for groups to provide advice on whether to build the power stations or not.

Identify groups who will want to contribute to the inquiry, including groups for and against the building of the nuclear power stations. Explain the arguments they may make, including any key scientific issues.

Figure 26 Study question 2 A181/2 Q6 (OCR, 2012h)



Question			Answer	Marks	Guidance
6			**Level 3: (5 – 6 marks)** Identifies a group affected in favour and suggests an appropriate argument and identifies a group against and suggests an appropriate argument. Answer should include reference to nuclear waste or ionising radiation. Quality of written communication does not impede communication of the science at this level. **Level 2: (3 – 4 marks)** Identifies groups affected, both in favour and against, and for at least one group suggests an appropriate argument. Some arguments may be given without identifying groups. Answer may include reference to radioactive materials. Quality of written communication partly impedes communication of the science at this level. **Level 1: (1 – 2 marks)** Suggests an appropriate argument for or against, may not have identified the group. Quality of written communication impedes communication of the science at this level. **Level 0: (0 marks)** Insufficient or irrelevant science. Answer not worthy of credit.	6	**This question is targeted at grades up to B** **Indicative scientific points may include:** **Groups in favour may include:** eg any environmental group. because: replaces pollution producing coal less CO_2 produced eg local workers / businesses because provides work during building and after eg health services / doctors because provides radioactive materials for diagnosis and treatment **Groups against may include:** eg people living in area nearby because they are afraid of being contaminated / irradiated eg conservative groups because it's not natural / it's new eg any environmental group because produces harmful radioactive waste waste very difficult to dispose of risk of accidents when waste being transported eg security services because may provide radioactive materials to terrorists **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**
			Total	**6**	

Figure 27 Study question 2 OCR Mark scheme A181/2 Q6 January 2012 (OCR, 2012f)

Study Mark scheme Question 2 Nuclear power station A181/02 Q6

Framework descriptor	**Mark scheme descriptors**	**Guidance**
B3 The argument or explanation of the claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence or argument (weak rebuttal).	**Level 3** Answer identifies group **for** and puts forward at least one piece of evidence for with scientific explanation. Answer identifies group **against** and puts forward at least one piece of evidence against with scientific explanation. Quality of written communication does not impede communication of the science at this level.	**This question is targeted at grades up to B.** Answers at level 3 must include reference radioactive materials / ionising radiation. (As candidates are required to put both sides of the argument, they are not expected to include any explicit rebuttals.) **Groups and arguments for nuclear power station** • environmental groups – reduces CO_2 emissions of power production – so reducing greenhouse gases; reduces particulate/acid rain gases – so reducing environmental damage • local people near old coal stations – less emissions from NPS so cleaner air; nuclear fuel much less bulky, so fewer lorries/rail trucks in and out • workers near PS – provides work during demolition / construction of PS **Groups and arguments against power station** • People living near NPS sites / People near nuclear waste disposal – concerns about ionising radiation during use / risk of accidents – radiation can cause cancer • environmental group – disposal of nuclear waste is a problem: ionising, long lasting
B2 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.	**Level 2** Answer identifies groups for and against NPS. Uses evidence for and against with some scientific reasoning for at least one argument. Quality of written communication partly impedes communication of the science at this level.	
B1 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning, (warrant) OR refers to limits of evidence.	**Level 1** Puts forward evidence for and against but may not make explicit links to groups. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Figure 28 Study question 2 Mark scheme derived from theoretical framework

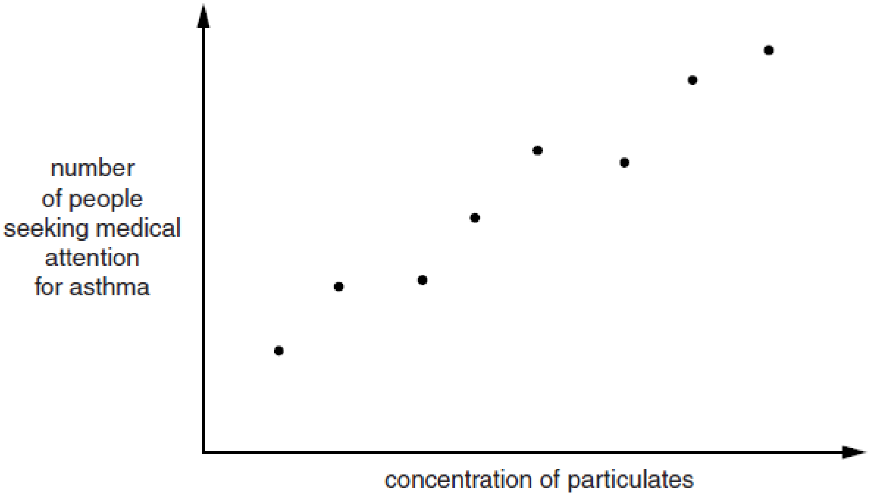
Study Question 3 Asthma A 171/01 Q2b, A171/2 Q1b

1 Scientists measure the concentration of particulates in the air in a town centre.

They do this on several days.

They also count the number of people seeking medical attention for asthma on the same days.

They plot their results on a scatter graph.



(b) A journalist talks to the scientists about their data before it is published.

The journalist writes a newspaper article using the scientists' data.

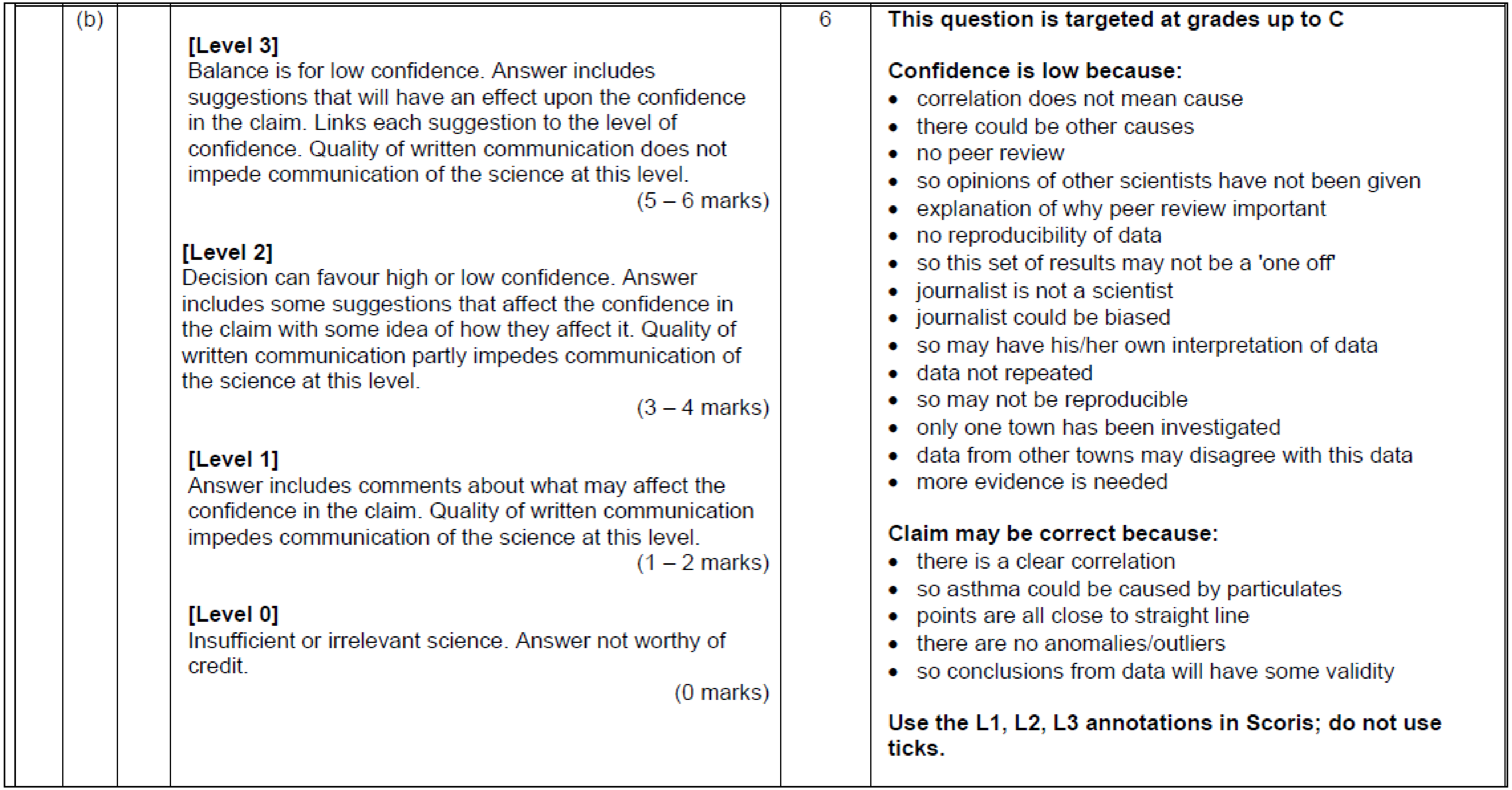
The article makes this claim.

> 'Asthma is caused by particulates in the air'.

How much confidence can be placed in the newspaper claim?

The quality of written communication will be assessed in your answer.

Figure 29 Study question 3 Asthma A 171/01 Q2b, A171/2 Q1b (OCR, 2012c)



	(b)		**[Level 3]** Balance is for low confidence. Answer includes suggestions that will have an effect upon the confidence in the claim. Links each suggestion to the level of confidence. Quality of written communication does not impede communication of the science at this level. (5 – 6 marks) **[Level 2]** Decision can favour high or low confidence. Answer includes some suggestions that affect the confidence in the claim with some idea of how they affect it. Quality of written communication partly impedes communication of the science at this level. (3 – 4 marks) **[Level 1]** Answer includes comments about what may affect the confidence in the claim. Quality of written communication impedes communication of the science at this level. (1 – 2 marks) **[Level 0]** Insufficient or irrelevant science. Answer not worthy of credit. (0 marks)	6	**This question is targeted at grades up to C** **Confidence is low because:** • correlation does not mean cause • there could be other causes • no peer review • so opinions of other scientists have not been given • explanation of why peer review important • no reproducibility of data • so this set of results may not be a 'one off' • journalist is not a scientist • journalist could be biased • so may have his/her own interpretation of data • data not repeated • so may not be reproducible • only one town has been investigated • data from other towns may disagree with this data • more evidence is needed **Claim may be correct because:** • there is a clear correlation • so asthma could be caused by particulates • points are all close to straight line • there are no anomalies/outliers • so conclusions from data will have some validity **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**

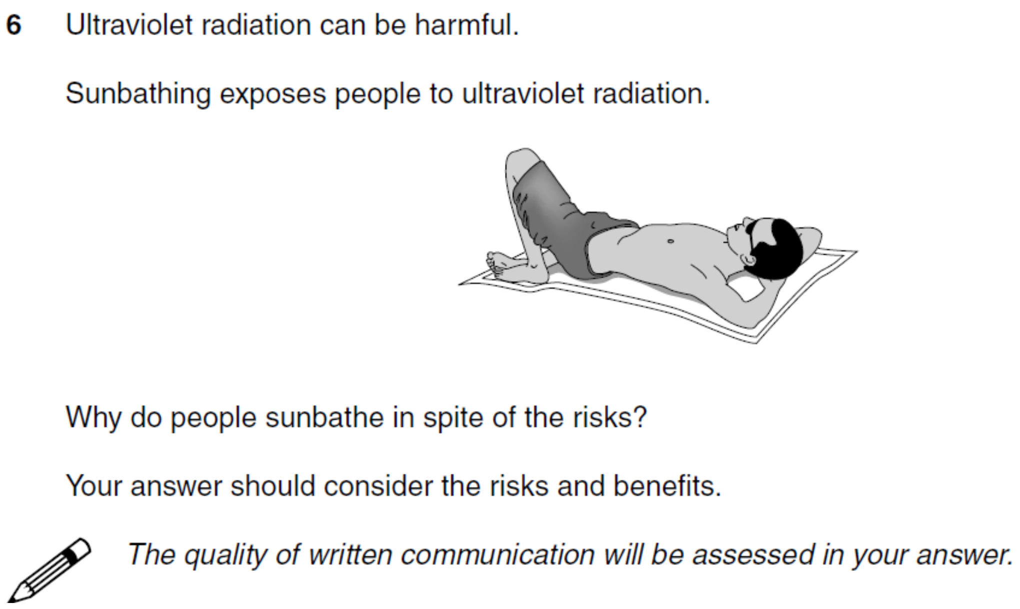
Figure 30 Study question 3 OCR Mark Scheme A171/01 Q2b and A171/2 Q1b (OCR, 2012i)

Study Mark scheme Question 3 Asthma A 171/01 Q2b, A171/2 Q1b

Framework descriptor	Mark scheme descriptors	Guidance
C3 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.	**Level 3** **Claim** Makes clear whether there can be confidence or not; uses evidence from the text to support claim; uses scientific reasoning / other scientific knowledge to support use of evidence; Gives a reason for uncertainty in claim / limits of evidence Quality of written communication does not impede communication of the science at this level.	This question is targeted at grades up to C. Candidates may argue for confidence in the claim of the journalist or lack of confidence in the claim. However for candidates to score level 3 with a 'confident' answer they will also need to include a rebuttal as the evidence in the text leads to a stronger 'no confidence' argument. **Evidence and supporting arguments for confidence** • data collected and analysed by scientist • there is a correlation between concentration of particulates and number of people • **but** there has been no scientific argument about how particulates cause asthma • **but** there could be some other emission alongside the particulates that **Evidence and supporting arguments for no confidence** • not yet published (in scientific journal) • so no (evidence of) peer review • **so** no other scientists have scrutinised the data • only carried out the experiment in one town • **so** not shown to be reproducible • **so** this could be a coincidence / some other factor which correlates with both • more data from other towns / times needed **to** increase confidence • journalist could be biased / have other reasons for making the statement • there are other known causes of asthma • no mechanism for causal link given by scientist / journalist
C2 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning, (warrant) / refers to limits of evidence.	**Level 2** **Claim** States whether there can be confidence or not. Provides some evidence from the text; Gives some other reasoning / refers to limits of evidence; Quality of written communication partly impedes communication of the science at this level.	
C1 May make clear the claim; Provides some relevant evidence or scientific reasoning. No reference to limitations of evidence or reasoning.	**Level 1** **Claim** May state whether there can be confidence or not; gives some supporting reasoning / develops an idea from text Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Figure 31 Study question 3 Mark scheme derived from theoretical framework

Study Question 4 Sunbathing A181/01 Q6, A181/02 Q4



6 Ultraviolet radiation can be harmful.

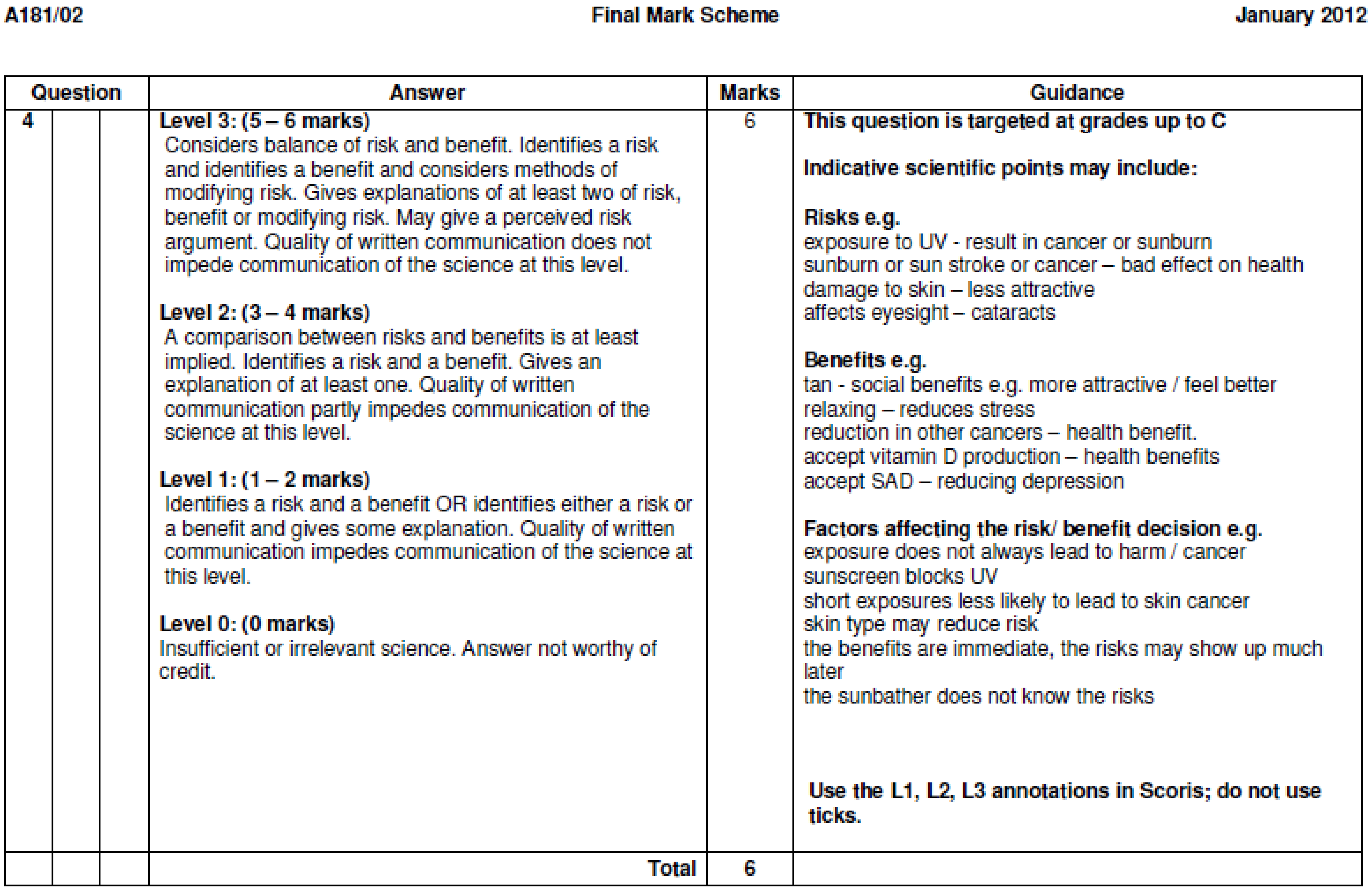
Sunbathing exposes people to ultraviolet radiation.

Why do people sunbathe in spite of the risks?

Your answer should consider the risks and benefits.

The quality of written communication will be assessed in your answer.

Figure 32 Study question 4 A181/01 Q6, A181/02 Q4 (OCR, 2012h)



A181/02 Final Mark Scheme January 2012

Question			Answer	Marks	Guidance
4			**Level 3: (5 – 6 marks)** Considers balance of risk and benefit. Identifies a risk and identifies a benefit and considers methods of modifying risk. Gives explanations of at least two of risk, benefit or modifying risk. May give a perceived risk argument. Quality of written communication does not impede communication of the science at this level. **Level 2: (3 – 4 marks)** A comparison between risks and benefits is at least implied. Identifies a risk and a benefit. Gives an explanation of at least one. Quality of written communication partly impedes communication of the science at this level. **Level 1: (1 – 2 marks)** Identifies a risk and a benefit OR identifies either a risk or a benefit and gives some explanation. Quality of written communication impedes communication of the science at this level. **Level 0: (0 marks)** Insufficient or irrelevant science. Answer not worthy of credit.	6	**This question is targeted at grades up to C** **Indicative scientific points may include:** **Risks e.g.** exposure to UV - result in cancer or sunburn sunburn or sun stroke or cancer – bad effect on health damage to skin – less attractive affects eyesight – cataracts **Benefits e.g.** tan - social benefits e.g. more attractive / feel better relaxing – reduces stress reduction in other cancers – health benefit. accept vitamin D production – health benefits accept SAD – reducing depression **Factors affecting the risk/ benefit decision e.g.** exposure does not always lead to harm / cancer sunscreen blocks UV short exposures less likely to lead to skin cancer skin type may reduce risk the benefits are immediate, the risks may show up much later the sunbather does not know the risks **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**
			Total	**6**	

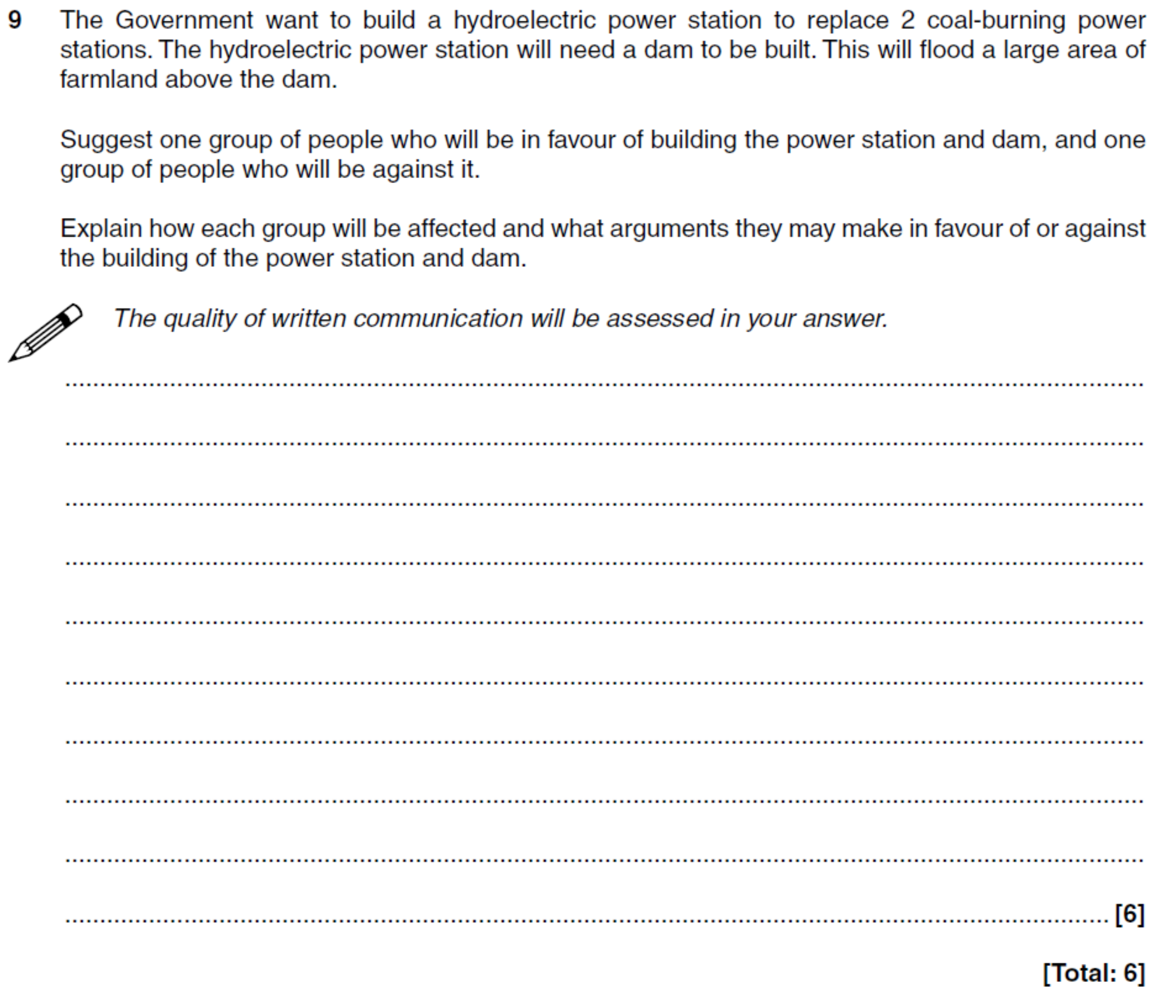
Figure 33 Study question 4 OCR Mark Scheme A181/01 Q6, A181/02 Q4 (OCR, 2012f)

Study Mark Scheme Question 4 Sunbathing A181/01 Q6, A181/02 Q4

Framework descriptor	Mark scheme descriptors	Guidance
C3 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.	**Level 3** Identifies a benefit and a risk and a method of modifying the risk or a reason for sunbathing in spite of risk (rebuttal). Gives some evidence or scientific reasoning for at least two of benefit, risk or modifying risk. Quality of written communication does not impede communication of the science at this level.	This question is targeted at grades up to **C.** The question asks for both risks and benefits of sunbathing to be considered. Note that the questions states that sunbathing exposes people to UV which can be harmful. As this is targeted up to **C**, Possible arguments **Benefits**: • tan – social benefits: feel healthier / more attractive /reduces stress • health benefits – vitamin D production, reduction in SAD **Risks:** • skin damage / sunburn • leading to skin cancer • cataracts • due to UV being ionising radiation • which damages cells **Reasons for sunbathing in spite of the risk:** • mitigate exposure to UV ; use sun cream /limit exposure time • exposure does not always cause harm / damage is not immediate so risk not perceived a high • sunbather may not know about the risks
C2 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning (warrant), OR refers to limits of evidence.	**Level 2** Identifies a benefit and a risk and gives some evidence or reasoning for at least one. Quality of written communication partly impedes communication of the science at this level.	
C1 May make clear the claim; Provides some relevant evidence or scientific reasoning. No reference to limitations of evidence or reasoning.	**Level 1** Identifies a risk **or** benefit and gives some evidence or reasoning for it. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Figure 34 Study question 4 Mark scheme derived from theoretical framework

Study Question 5 Hydroelectric power station A181/01 Q9



9 The Government want to build a hydroelectric power station to replace 2 coal-burning power stations. The hydroelectric power station will need a dam to be built. This will flood a large area of farmland above the dam.

Suggest one group of people who will be in favour of building the power station and dam, and one group of people who will be against it.

Explain how each group will be affected and what arguments they may make in favour of or against the building of the power station and dam.

The quality of written communication will be assessed in your answer.

....

....

....

....

....

....

....

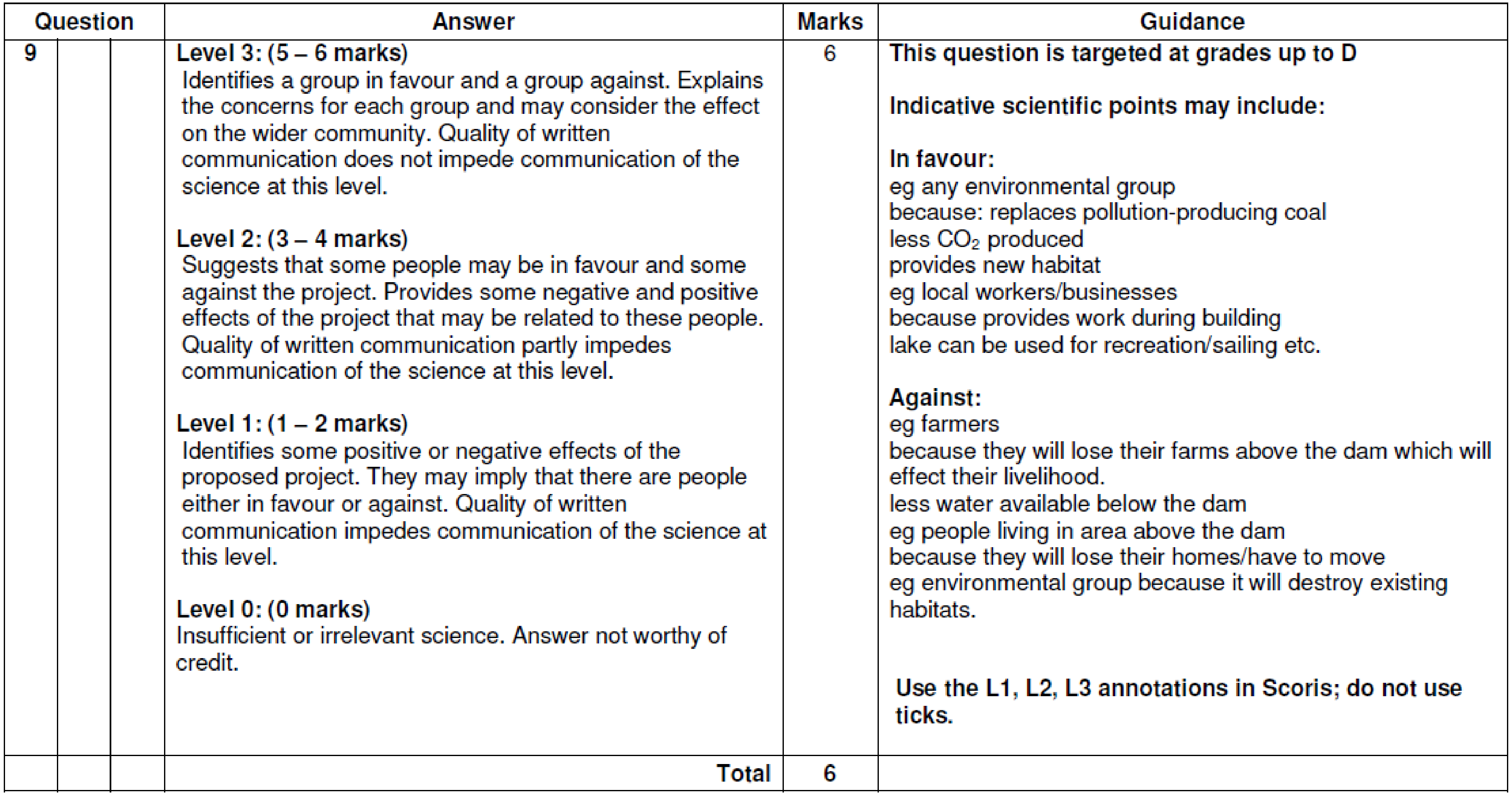
....

....

.... **[6]**

[Total: 6]

Figure 35 Study question 5 A181/01 Q9 (OCR, 2012g)



Question			Answer	Marks	Guidance
9			**Level 3: (5 – 6 marks)** Identifies a group in favour and a group against. Explains the concerns for each group and may consider the effect on the wider community. Quality of written communication does not impede communication of the science at this level. **Level 2: (3 – 4 marks)** Suggests that some people may be in favour and some against the project. Provides some negative and positive effects of the project that may be related to these people. Quality of written communication partly impedes communication of the science at this level. **Level 1: (1 – 2 marks)** Identifies some positive or negative effects of the proposed project. They may imply that there are people either in favour or against. Quality of written communication impedes communication of the science at this level. **Level 0: (0 marks)** Insufficient or irrelevant science. Answer not worthy of credit.	6	**This question is targeted at grades up to D** **Indicative scientific points may include:** **In favour:** eg any environmental group because: replaces pollution-producing coal less CO_2 produced provides new habitat eg local workers/businesses because provides work during building lake can be used for recreation/sailing etc. **Against:** eg farmers because they will lose their farms above the dam which will effect their livelihood. less water available below the dam eg people living in area above the dam because they will lose their homes/have to move eg environmental group because it will destroy existing habitats. **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**
			Total	6	

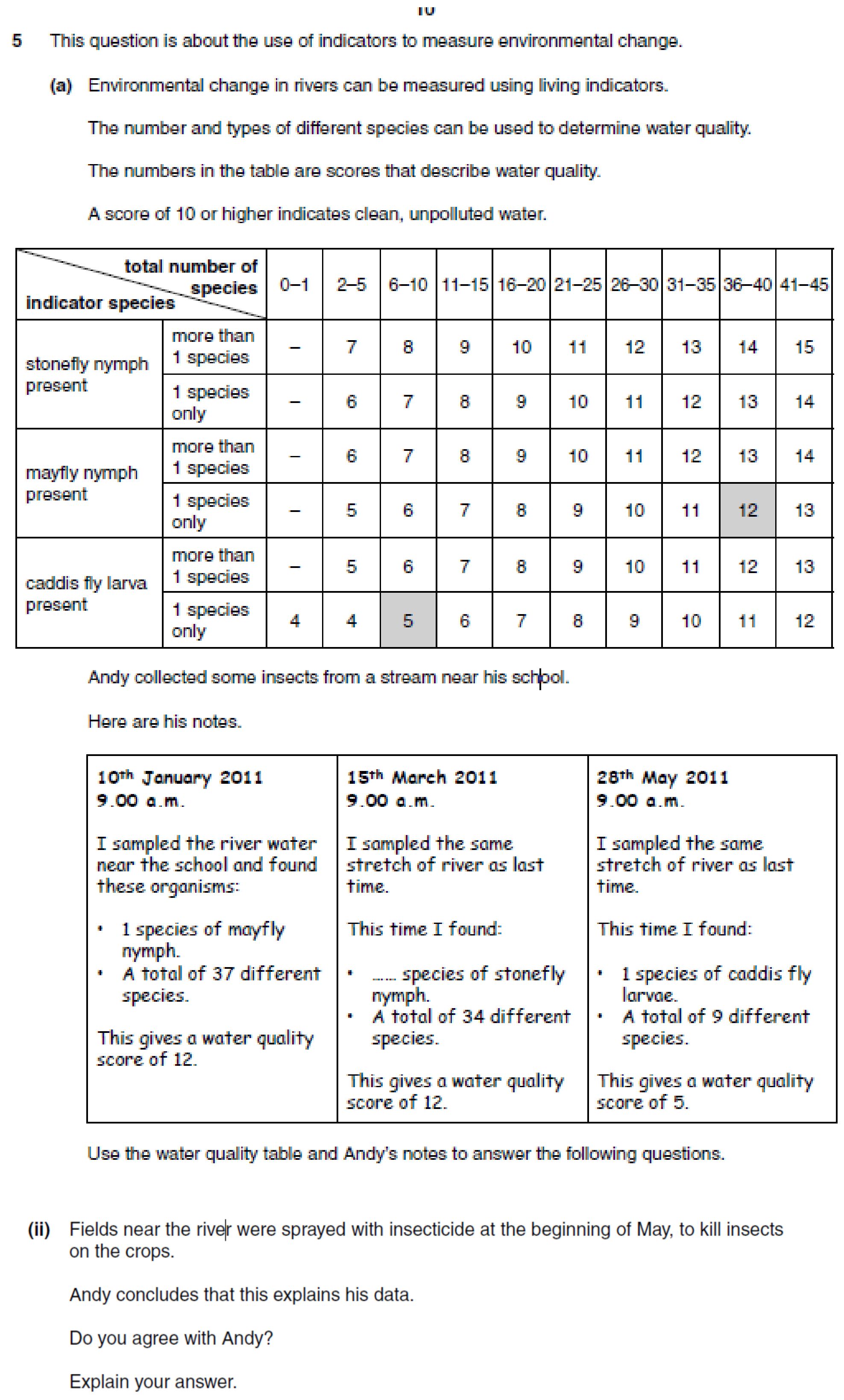
Figure 36 Study question 5 OCR Mark Scheme A181/01 Q9 (OCR, 2012e)

Study Mark Scheme Question 5 Hydroelectric power station A181/01 Q9

Framework descriptor	**Mark scheme descriptors**	**Guidance**
D3 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning (warrant), OR refers to limits of evidence.	**Level 3** Answer identifies group **for** and puts forward at least one piece of evidence Answer identifies group **against** and puts forward at least one piece of evidence Answers at this level must include some scientific reasoning or refer to a limit on evidence / counter argument for one group. Quality of written communication does not impede communication of the science at this level.	This question is targeted at grades up to D. Possible marking points **Groups and arguments for hydroelectric power station** • environmental groups – reduces CO_2 emissions of power production – so reducing greenhouse gases; reduces particulate/acid rain gases – so reducing environmental damage • local people near old coal stations – cleaner air; • workers near HEPS – provides work during construction of dam / operationally / tourism **Groups and arguments against hydroelectric power station** • farmers whose land will be flooded – loss of income/jobs/livelihood • people living in flooded area above dam – have to move home • environmental group – loss of habitats
D2 May make clear the claim; Provides some relevant evidence or scientific reasoning. No reference to limitations of evidence or reasoning.	**Level 2** Answer identifies groups for and against ; provides a reason for each group Quality of written communication partly impedes communication of the science at this level.	
D1 Identifies some relevant factor, evidence or reasoning but the links are weak	**Level 1** Puts forward some reasons for or against but may not link to groups. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Figure 37 Study question 5 Mark scheme derived from theoretical framework

Study Question 6 Environmental indicators A161/01 Q5



10

5 This question is about the use of indicators to measure environmental change.

(a) Environmental change in rivers can be measured using living indicators.

The number and types of different species can be used to determine water quality.

The numbers in the table are scores that describe water quality.

A score of 10 or higher indicates clean, unpolluted water.

indicator species	**total number of species**	0–1	2–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45
stonefly nymph present	more than 1 species	–	7	8	9	10	11	12	13	14	15
	1 species only	–	6	7	8	9	10	11	12	13	14
mayfly nymph present	more than 1 species	–	6	7	8	9	10	11	12	13	14
	1 species only	–	5	6	7	8	9	10	11	12	13
caddis fly larva present	more than 1 species	–	5	6	7	8	9	10	11	12	13
	1 species only	4	4	5	6	7	8	9	10	11	12

Andy collected some insects from a stream near his school.

Here are his notes.

10th January 2011 **9.00 a.m.**	**15th March 2011** **9.00 a.m.**	**28th May 2011** **9.00 a.m.**
I sampled the river water near the school and found these organisms: • 1 species of mayfly nymph. • A total of 37 different species. This gives a water quality score of 12.	I sampled the same stretch of river as last time. This time I found: • ...... species of stonefly nymph. • A total of 34 different species. This gives a water quality score of 12.	I sampled the same stretch of river as last time. This time I found: • 1 species of caddis fly larvae. • A total of 9 different species. This gives a water quality score of 5.

Use the water quality table and Andy's notes to answer the following questions.

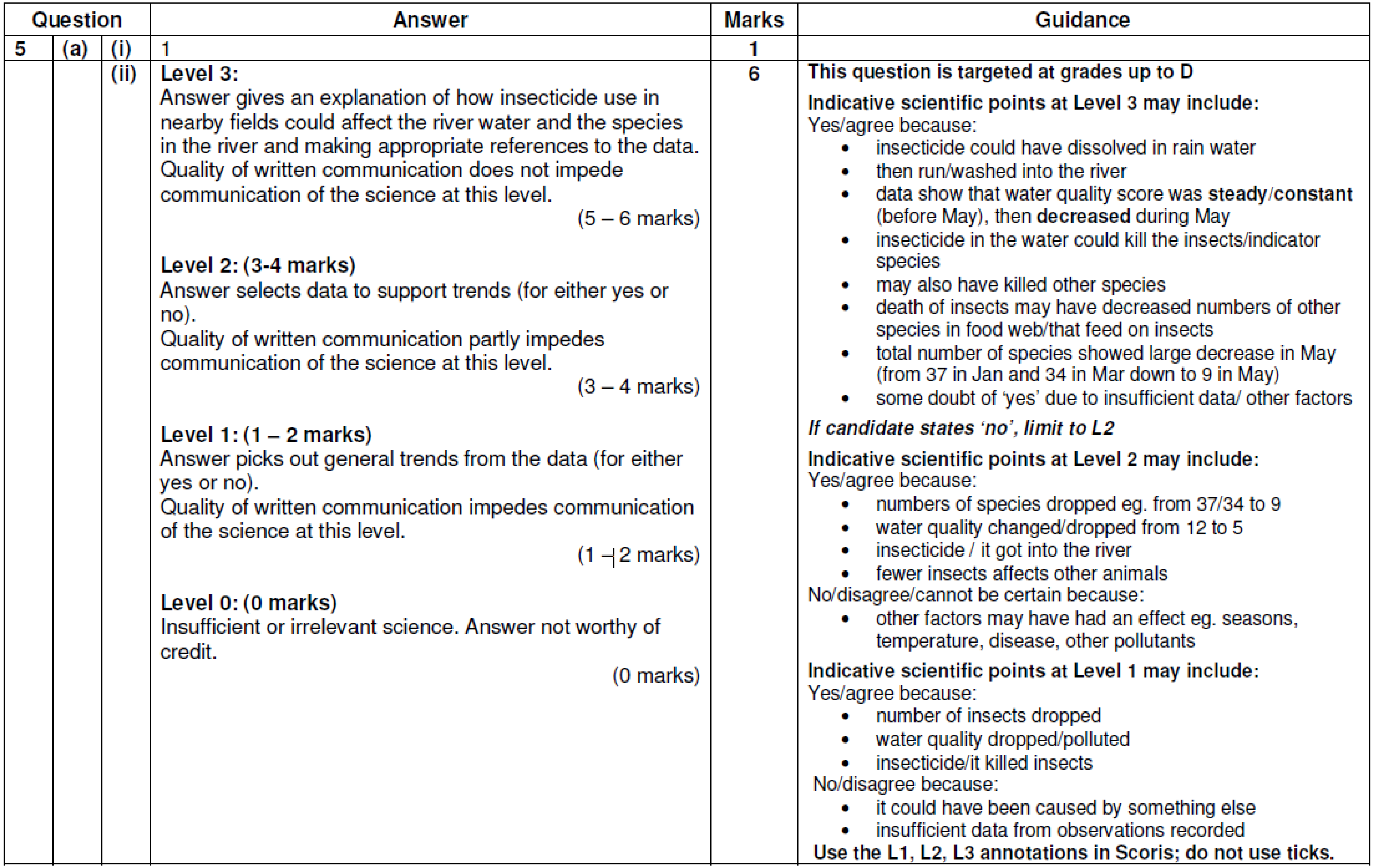
(ii) Fields near the river were sprayed with insecticide at the beginning of May, to kill insects on the crops.

Andy concludes that this explains his data.

Do you agree with Andy?

Explain your answer.

Figure 38 Question 6 A161/01 Q5 (OCR, 2012a)



Question			Answer	Marks	Guidance
5	(a)	(i)	1	1	
		(ii)	**Level 3:** Answer gives an explanation of how insecticide use in nearby fields could affect the river water and the species in the river and making appropriate references to the data. Quality of written communication does not impede communication of the science at this level. (5 – 6 marks) **Level 2: (3-4 marks)** Answer selects data to support trends (for either yes or no). Quality of written communication partly impedes communication of the science at this level. (3 – 4 marks) **Level 1: (1 – 2 marks)** Answer picks out general trends from the data (for either yes or no). Quality of written communication impedes communication of the science at this level. (1 –2 marks) **Level 0: (0 marks)** Insufficient or irrelevant science. Answer not worthy of credit. (0 marks)	6	**This question is targeted at grades up to D** **Indicative scientific points at Level 3 may include:** Yes/agree because: • insecticide could have dissolved in rain water • then run/washed into the river • data show that water quality score was **steady**/**constant** (before May), then **decreased** during May • insecticide in the water could kill the insects/indicator species • may also have killed other species • death of insects may have decreased numbers of other species in food web/that feed on insects • total number of species showed large decrease in May (from 37 in Jan and 34 in Mar down to 9 in May) • some doubt of 'yes' due to insufficient data/ other factors ***If candidate states 'no', limit to L2*** **Indicative scientific points at Level 2 may include:** Yes/agree because: • numbers of species dropped eg. from 37/34 to 9 • water quality changed/dropped from 12 to 5 • insecticide / it got into the river • fewer insects affects other animals No/disagree/cannot be certain because: • other factors may have had an effect eg. seasons, temperature, disease, other pollutants **Indicative scientific points at Level 1 may include:** Yes/agree because: • number of insects dropped • water quality dropped/polluted • insecticide/it killed insects No/disagree because: • it could have been caused by something else • insufficient data from observations recorded **Use the L1, L2, L3 annotations in Scoris; do not use ticks.**

Figure 39 OCR Mark Scheme A161/01 Q5 (OCR, 2012b)

Appendix 2 OCR GCSE Science A January 2012 – Facility values

Table 15 shows item level data for all questions with levels-based mark schemes in the January 2012 series of papers for the OCR GCSE Science A *Twenty First Century Science* suite:

- OCR GCSE Science A (B1 C1 P1) A141/01, A141/02, and GCSE Science (B2 C2 P2) A142/01, A142/02. (CA, 2012a, 2012b, 2012c, 2012d; OCR, 2012m, 2012n, 2012o, 2012p)
- OCR GCSE Biology A (B1 B2 B3) A161/01, A161/02 (CA, 2012e, 2012f; OCR, 2012e, 2012f)
- OCR GCSE Chemistry A (C1 C2 C3) A171/01,A171/02 (CA, 2012g, 2012h; OCR, 2012g, 2012j)
- OCR GCSE Physics A (P1 P2 P3) A181/01, A181/02 (CA, 2012i, 2012j; OCR, 2012k, 2012l)

Foundation tier papers target questions up to grade C and all have paper numbers ending /01.

Higher tier papers target grades D-A* and all paper numbers end /02.

Facility of a question is the mean mark awarded for the question as a proportion of the maximum mark for the question. Facility of a question at a grade is calculated using the mean mark for the questions for candidates who achieved that grade on the paper (Elliott & Johnson, 2007).

Target grade is the grade identified on the mark scheme as the intended demand of the question and mark scheme.

Table 15 Item level data for OCR GCSE Science A January 2012

Paper	Question	Target grade	Facility at target grade	Facility at target grade F tier	Facility at target grade H Tier
181/02	2	A/A*	0.62/0.78		0.70†
161/02	4ai	A/A*	0.62/0.77		0.69†
141/02	5b	A/A*	0.55/0.77		0.66†
161/02	6	A/A*	0.49/0.78		0.64†
142/02	7a	A/A*	0.46/0.72		0.59†
142/02	10	A/A*	0.39/0.64		0.52†
171/02	2a	A/A*	0.44/0.54		0.49†
141/02	8	A	0.63		0.63
171/02	3ci	A	0.42		0.42
181/02	6	B	0.55		0.55
142/01	2a#	C	0.65	0.65	
142/02	3a#	C	0.48		0.48
181/01	6#	C	0.65	0.65	
181/02	4#	C	0.57		0.57
141/01	3a#	C	0.55	0.55	
141/02	3a#	C	0.44		0.44
161/01	2c#	C	0.52	0.52	
161/02	2b#	C	0.49		0.49
171/01	2b#	C	0.35	0.35	
171/02	1b#	C	0.27		0.27
142/01	10	C	0.68	0.68	
141/01	8	C	0.59	0.59	
171/01	3d	C	0.24	0.24	
181/01	9	D	0.58	0.58	
161/01	5aii	D	0.21	0.21	
141/01	4d	E	0.35	0.35	
161/01	4c	E	0.25	0.25	
171/01	5a	E	0.15	0.15	
142/01	7	E	0.12	0.12	
181/01	2	F	0.26	0.26	
Mean			**0.48**	**0.41**	**0.54**
Standard deviation			**0.17**	**0.20**	**0.12**

†The mean of the facilities for the question at A and A*.

Pairs of questions marked #, are identical questions included in both the foundation tier (01) and higher tier papers (02).

The shaded rows show the questions used in the study.

Appendix 3 Comparison of marks awarded using the theoretical framework with those awarded by the original examiners

Terms used throughout this appendix

Examiner mark (EM) the mark awarded by the examiner when originally marked using the OCR mark scheme.

Facility of a question is the mean mark awarded for the question as a proportion of the maximum mark for the question. Facility of a question at a grade is calculated using the mean mark for the questions for candidates who achieved that grade on the paper. (Elliott & Johnson, 2007)

Item level data (ILD) provides information about the facility of the question and how it varies for different ability candidates (see section 2.6).

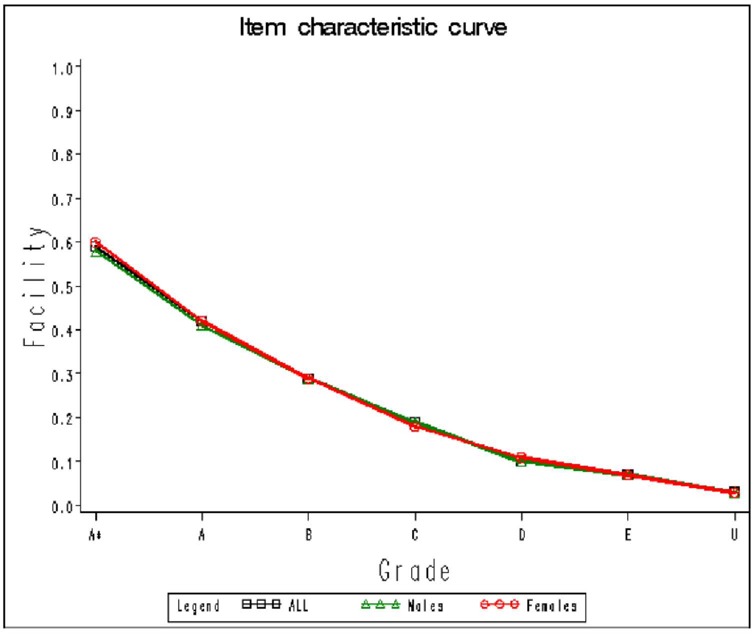
P_{Agr1} Proportion of answers for which two markers were within agreement to within one mark.

Researcher mark (RM) mark awarded for question by the researcher using mark scheme derived from theoretical framework

Appendix 3.1 Question 1 Polymers – A171/2 Q3(c)(i)

Item-level data for Question 1 Polymers – A171/2 Q3(c)(i) (CA, 2012j)

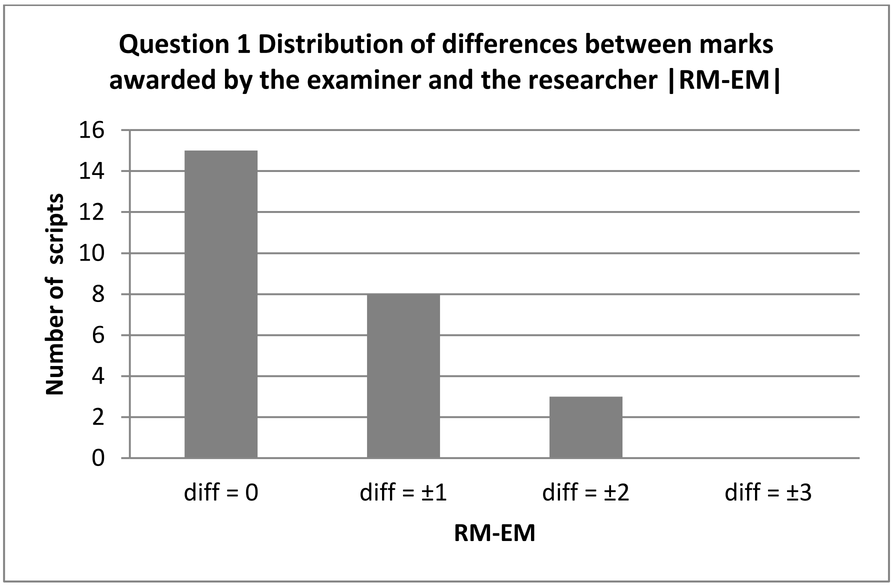
20497 candidates	**facility (cohort)**	**facility at each grade**								
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
A171/2 Q3ci	0.33	0.59	0.42	0.29	0.19	0.10	0.07	-	-	0.03



Raw data from research

Question 1 Polymers		**mark awarded using original mark scheme**	**mark awarded for question using mark scheme derived from theoretical framework**	**difference between marks awarded**
candidate code	**script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM - EM**
1710201	B	3	2	-1
1710202	Y	4	4	0
1710203	U	0	0	0
1710204	H	6	6	0
1710205	A	2	2	0
1710206	C	3	3	0
1710207	J	3	3	0
1710208	G	3	4	1
1710209	T	6	4	-2
1710210	V	0	0	0
1710211	S	6	5	-1
1710212	P	0	1	1
1710213	W	0	0	0
1710214	K	2	0	-2
1710215	R	2	2	0
1710216	M	4	2	-2
1710217	Q	4	5	1
1710218	O	5	5	0
1710219	N	2	2	0
1710220	L	4	4	0
1710221	F	2	1	-1
1710222	X	1	0	-1
1710223	D	5	5	0
1710224	I	1	0	-1
1710225	Z	2	2	0
1710226	E	2	2	0

Reliability indicators (see section 4.3.4)



Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-EM)	Mean absolute difference
Q1 Polymers	26	6	0.88	-0.31	0.54

Kendall's tau correlation

Question 1 Polymers		Examiner mark	Researcher mark
Examiner mark	correlation coefficient Significance	1	0.827 0.0000001
Researcher mark	correlation coefficient Significance	0.827 0.000	1

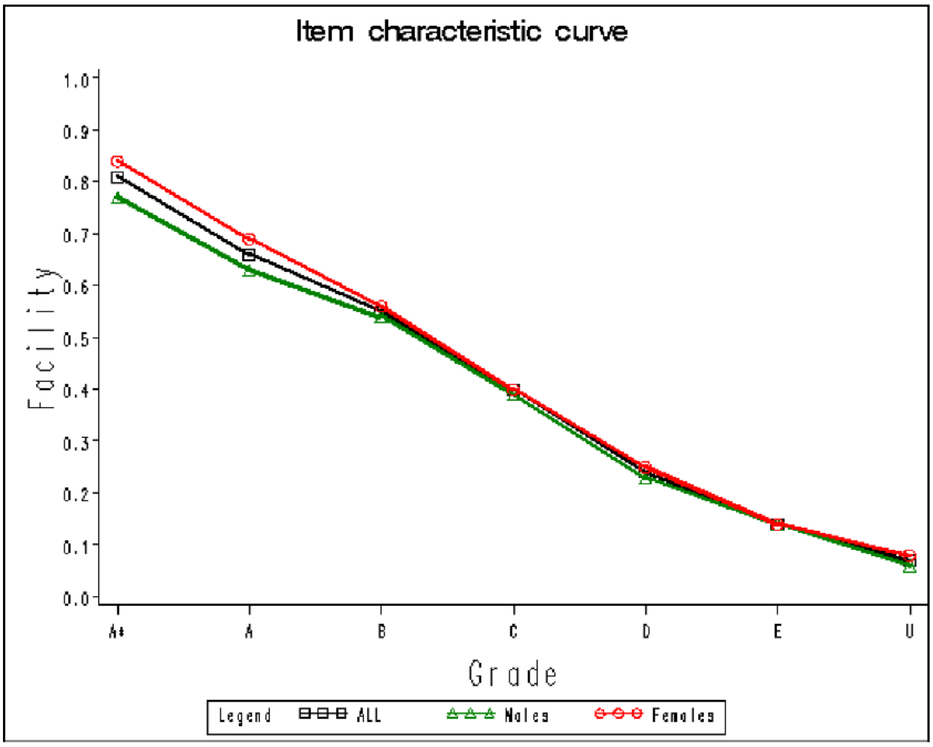
N=26

Shaded boxes show those correlations that are significant at the 0.01 level (2-tailed).

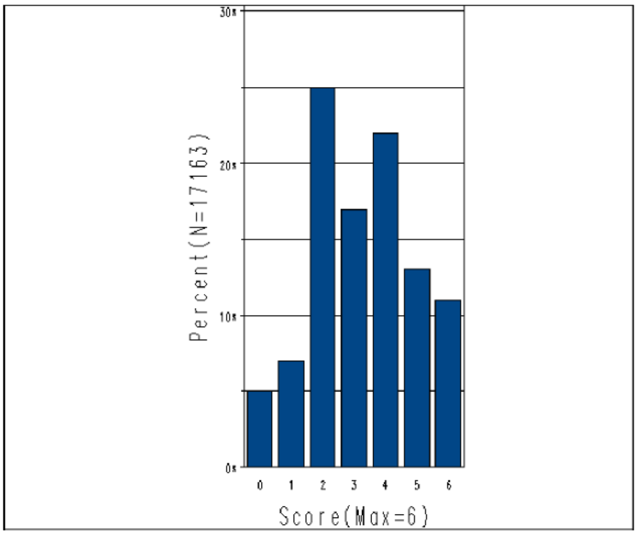
Appendix 3.2 Question 2 Nuclear power station – A181/2 Q6

Examination statistics for the question for whole cohort (CA, 2012j)

17163 candidates	**facility (cohort)**	**facility at each grade**								
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
A181/2 Q6	0.54	0.81	0.66	0.55	0.40	0.24	0.14	-	-	0.07



Item Frequency Distribution

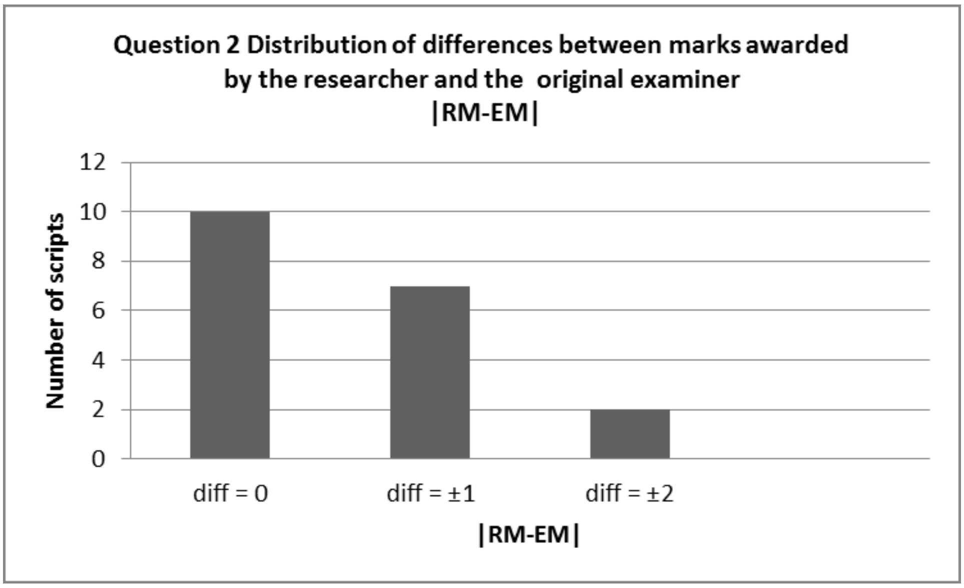


Raw data from research

Question 2 Nuclear power station		mark awarded using original mark scheme	mark awarded for question using mark scheme derived from theoretical framework	difference between marks awarded
Candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM – EM**
1810201	B	2	2	0
1810202	D	2	2	0
1810203	F	2	2	0
1810204	H	2	2	0
1810205	A	3	2	-1
1810206	C	3	3	0
1810207	J	3	2	-1
1810208	E	3	4	1
1810209	T	4	4	0
1810210	G	4	3	-1
1810211	S	4	3	-1
1810212	P	4	3	-1
1810213	I	5	3	-2
1810214	K	5	5	0
1810215	R	5	5	0
1810216	M	5	5	0
1810217	Q	6	5	-1
1810218	O	6	6	0
1810219	N	6	4	-2
1810220*	L	6	2	-4

*On inspection, 6 marks should not have been awarded to candidate L on the original mark scheme. This data was not included in any statistical analysis. The script for Candidate Lis shown in figure 5.4 in Chapter 5.

Reliability indicators (see section 4.3.4)



Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-EM)	Mean absolute difference
Q2 Nuclear power station	19	6	0.89	-0.53	0.63

Kendall's tau correlation

Question 2 Nuclear power station		Examiner mark	Researcher mark
Examiner mark	correlation coefficient Significance	1	0.751 0.0001
Researcher mark	correlation coefficient Significance	0.751 0.0001	1

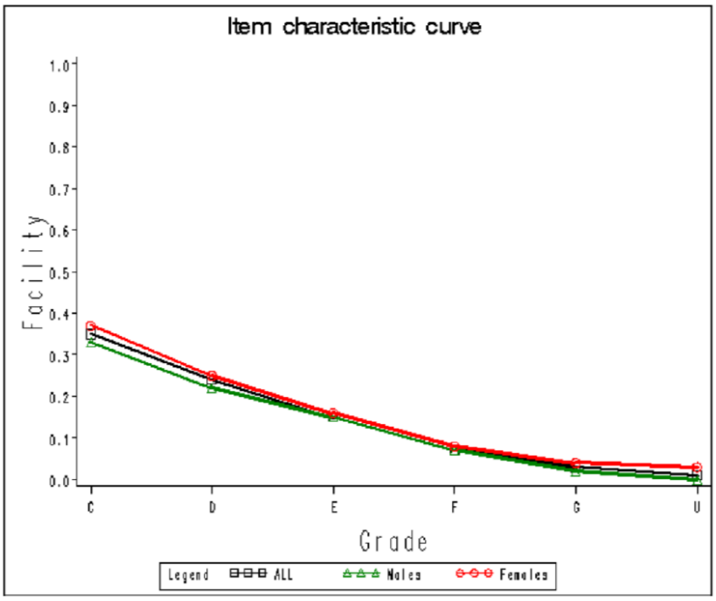
N=19

Shaded boxes show the correlation is significant at the 0.01 level (2-tailed).

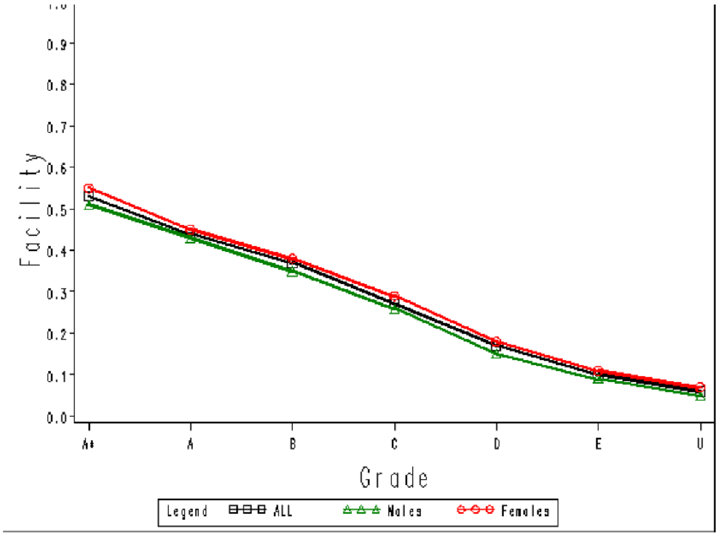
Appendix 3.3 Question 3 Asthma – A 171/01 Q2b, A171/2 Q1b

Examination statistics for question for whole cohort(CA, 2012g, 2012h)

	facility (cohort)	**facility at each grade**								
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
Foundation A171/01 2b (2789 candidates)	0.22	-	-	-	0.35	0.24	0.15	0.07	0.03	0.01
Higher A171/02 1b (20497 candidates)	0.37	0.53	0.44	0.37	0.27	0.17	0.10	-	-	0.06



Foundation Tier



Higher Tier

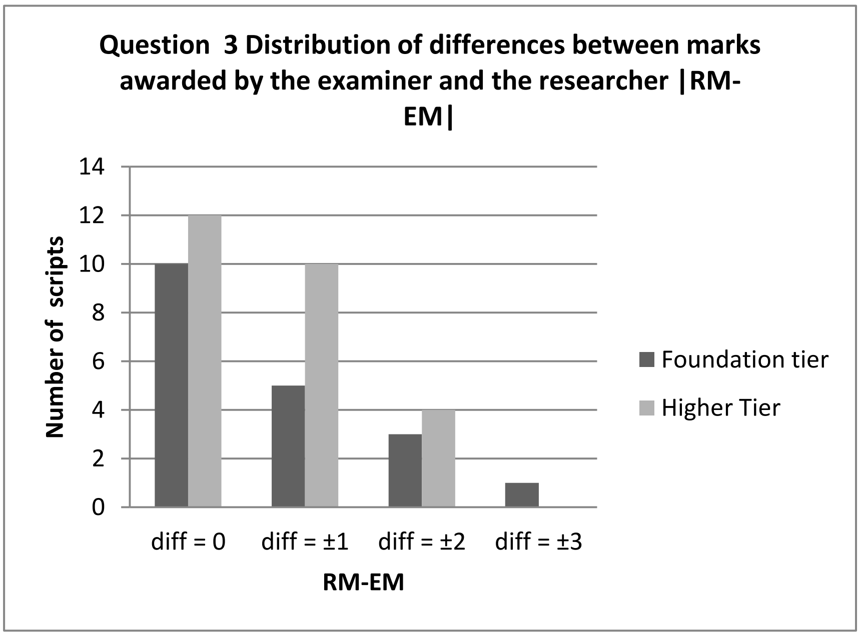
Raw data from research

Foundation Tier Q3 Asthma				
candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM - EM**
1710101	B	2	2	0
1710102	D	2	1	-1
1710103	F	2	1	-1
1710104	H	2	2	0
1710105	A	3	1	-2
1710106	C	3	4	1
1710107	J	3	4	1
1710108	E	3	1	-2
1710109	T	4	4	0
1710110	G	4	4	0
1710111	S	4	4	0
1710112	P	4	4	0
1710113	I	5	4	-1
1710114	K	5	5	0
1710115	R	5	5	0
1710116	M	5	5	0
1710117	Q	6	3	-3
1710118	O	6	6	0
1710119	N	6	4	-2

Higher Tier Q3 Asthma				
candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM - EM**
1710201	B	2	2	0
1710202	Y	2	3	1
1710203	U	2	2	0
1710204	H	2	3	1
1710205	A	3	3	0
1710206	C	3	5	2
1710207	J	3	4	1
1710208	G	3	4	1
1710209	T	4	4	0
1710210	V	4	2	-2
1710211	S	4	4	0
1710212	P	4	4	0
1710213	W	5	3	-2
1710214	K	5	4	-1
1710215	R	5	5	0
1710216	M	5	3	-2
1710217	Q	6	6	0
1710218	O	6	6	0
1710219	N	6	6	0
1710220	L	6	6	0
1710221	F	2	1	-1
1710222	X	4	4	0
1710223	D	1	2	1
1710224	I	3	4	1
1710225	Z	3	4	1
1710226	E	6	5	-1

Reliability indicators (see section 4.3.4)

Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-EM)	Mean absolute difference
Q3 Asthma Foundation Tier	19	6	0.79	-0.53	0.74
Q3 Asthma Higher Tier	26	6	0.85	0.00	0.71



Kendall's tau correlations

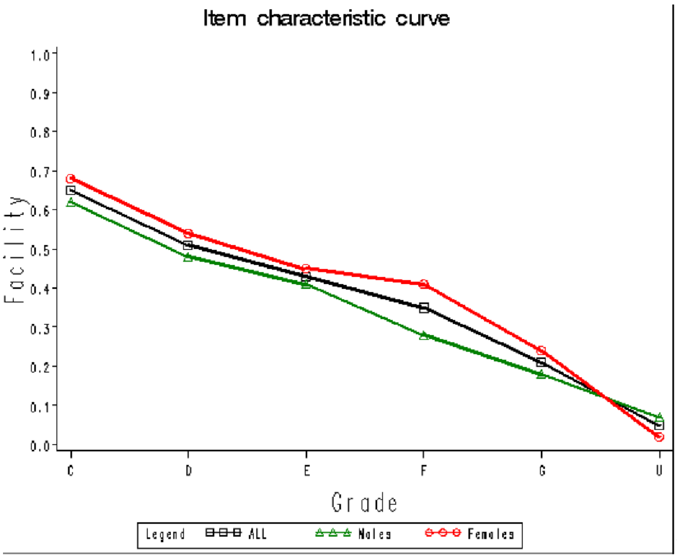
Question 3 Asthma – Foundation Tier *N=19*		Examiner mark	Researcher mark
Examiner mark	correlation coefficient Significance	1	0.614 0.001
Researcher mark	correlation coefficient Significance	0.614 0.001	1
Question 3 Asthma – Higher Tier *N=26*		**Examiner mark**	**Researcher mark**
Examiner mark	correlation coefficient Significance	1	0.632 0.000
Researcher mark	correlation coefficient Significance	0.632 0.000	1

Shaded boxes show the correlation is significant at the 0.01 level (2-tailed).

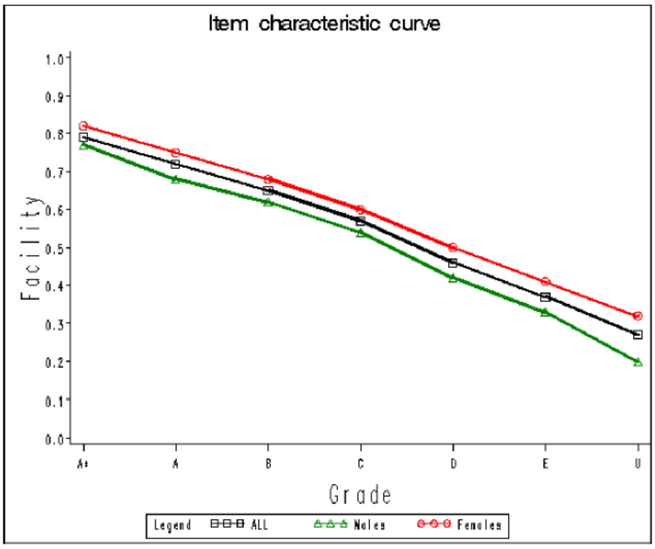
Appendix 3.4 Q4 Sunbathing A181/01 Q6, A181/02 Q4

Examination statistics for whole cohort (CA, 2012i, 2012j)

	facility (cohort)	**facility at each grade**								
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**
Foundation **A181/01 Q6** (2465 candidates)	0.50				0.65	0.51	0.43	0.35	0.21	0.05
Higher **A181/02 Q4** (17163 candidates)	0.37	0.79	0.72	0.65	0.57	0.46	0.37			0.27



Foundation Tier



Higher Tier

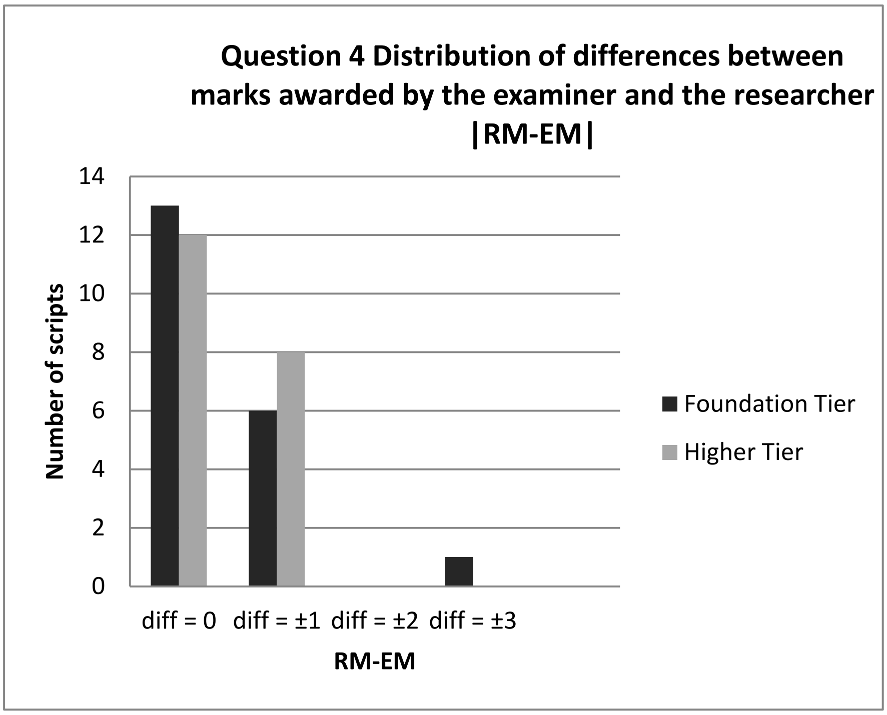
Raw data from research

Foundation Tier Q4 Sunbathing				
candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM - EM**
1810101	B	2	1	-1
1810102	D	2	1	-1
1810103	F	4	5	1
1810104	H	3	3	0
1810105	A	4	3	-1
1810106	C	0	0	0
1810107	J	2	2	0
1810108	E	2	2	0
1810109	T	5	5	0
1810110	G	6	6	0
1810111	S	3	3	0
1810112	P	1	1	0
1810113	I	4	5	1
1810114	K	2	2	0
1810115	R	3	3	0
1810116	M	2	2	0
1810117	Q	4	3	-1
1810118	O	4	4	0
1810119	N	6	3	-3
1810120	L	4	4	0

Higher Tier Q4 Sunbathing				
candidate code	**Script code**	**examiner mark (EM)**	**researcher mark (RM)**	**RM - EM**
1810201	B	6	5	-1
1810202	D	2	3	1
1810203	F	4	4	0
1810204	H	3	3	0
1810205	A	3	4	1
1810206	C	4	4	0
1810207	J	3	2	-1
1810208	E	4	4	0
1810209	T	5	4	-1
1810210	G	3	3	0
1810211	S	4	4	0
1810212	P	2	3	1
1810213	I	3	4	1
1810214	K	1	1	0
1810215	R	5	5	0
1810216	M	4	3	-1
1810217	Q	4	4	0
1810218	O	6	6	0
1810219	N	2	2	0
1810220	L	4	4	0

Reliability indicators (see section 4.3.4)

Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-EM)	Mean absolute difference
Q4 Sunbathing Foundation Tier	20	6	0.95	-0.25	0.45
Q4 Sunbathing Higher Tier	26	6	1.00	0.00	0.42



Kendall's tau correlations

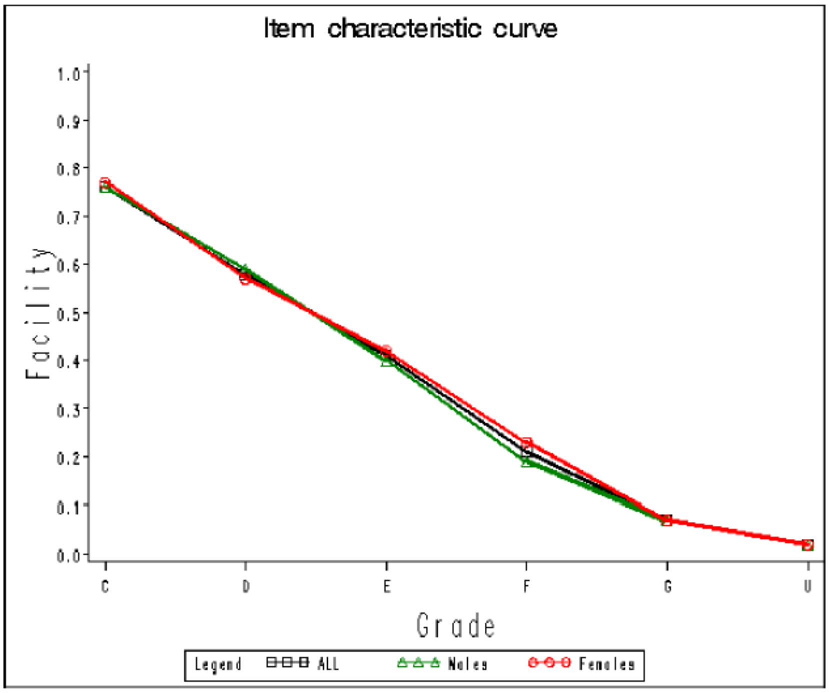
Question 4 Sunbathing – Foundation Tier *N=20*		Examiner mark	Researcher mark
Examiner mark	correlation coefficient Significance	1	0.824 0.000
Researcher mark	correlation coefficient Significance	0.824 0.000	1
Question 4 Sunbathing – Higher Tier *N=20*		**Examiner mark**	**Researcher mark**
Examiner mark	correlation coefficient Significance	1	0.784 0.000
Researcher mark	correlation coefficient Significance	0.784 0.000	1

Shaded boxes show the correlation is significant at the 0.01 level (2-tailed).

Appendix 3.5 Question 5 Power station – A 181/01 Q9

Examination statistics for question for whole cohort (CA, 2012i)

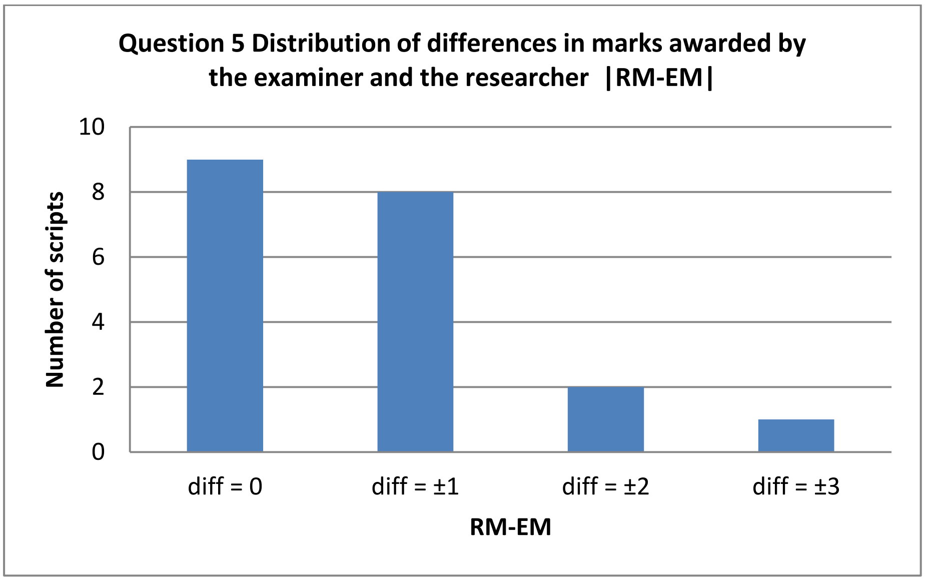
	facility (cohort)	facility at each grade								
		A*	A	B	C	D	E	F	G	U
A 181/01 Q9 (2465 candidates)	0.54	-	-	-	0.76	0.58	0.41	0.21	0.07	0.02



Raw data from research

Question 5 Power station		mark awarded using original mark scheme	mark awarded for question using mark scheme derived from theoretical framework	difference between marks awarded
Candidate code	Script code	examiner mark (EM)	researcher mark (RM)	RM - EM
1810101	B	2	2	0
1810102	D	2	2	0
1810103	F	2	1	-1
1810104	H	2	2	0
1810105	A	3	3	0
1810106	C	3	3	0
1810107	J	3	2	-1
1810108	E	3	3	0
1810109	T	4	3	-1
1810110	G	4	3	-1
1810111	S	4	3	-1
1810112	P	4	4	0
1810113	I	5	3	-2
1810114	K	5	3	-2
1810115	R	5	6	1
1810116	M	5	4	-1
1810117	Q	6	6	0
1810118	O	6	6	0
1810119	N	6	3	-3
1810120	L	6	5	-1

Reliability indicators (see section 4.3.4)



Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-EM)	Mean absolute difference
Q5 Power station	20	6	0.85	-0.65	0.75

Kendall’s tau correlation

Question 5 Power station		Examiner mark	Researcher mark
Examiner mark	correlation coefficient	1	0.744
	Significance		0.000
Researcher mark	correlation coefficient	0.744	1
	Significance	0.000	

N=20

Shaded boxes show the correlation is significant at the 0.01 level (2-tailed).

Appendix 4 Comparison of marks awarded during the validation of the framework

Terms used throughout this appendix

P_{Agr1} Proportion of answers for which two markers were within agreement to within one mark.

Principal Examiner mark (PM) mark awarded for question by the Principal Examiner (PE) using mark scheme derived from theoretical framework.

Researcher mark (RM) mark awarded for question by the researcher using mark scheme derived from theoretical framework

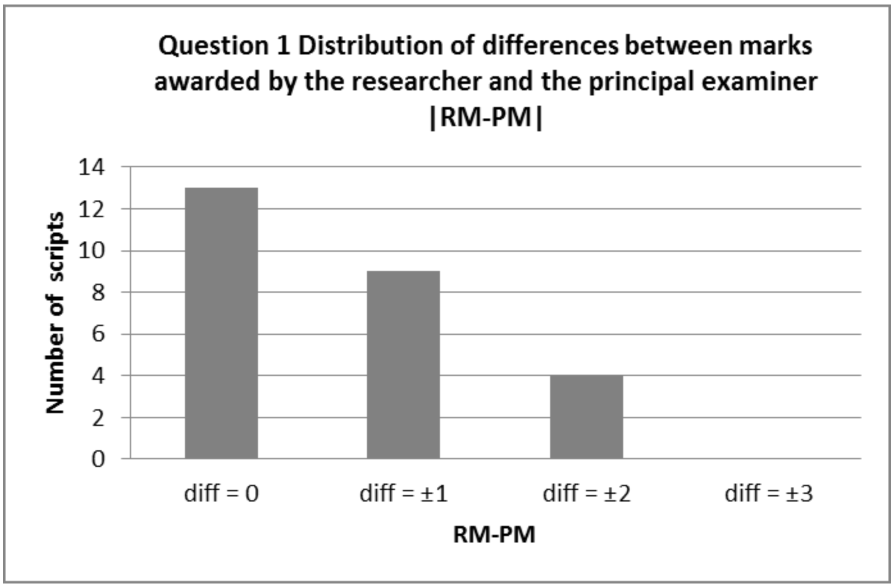
Standardised mark (SM) – the mark awarded using the mark scheme derived from theoretical framework as agreed by the researcher and PEs after a standardisation discussion. (See section 2.4 for more about standardisation of mark schemes.)

Validation : Question 1 Polymers – A171/2 Q3(c)(i)

Raw data from research

Question 1 Polymers		**mark awarded for question using mark scheme derived from theoretical framework**		**differences in marks awarded**
candidate code	**script code**	**researcher mark (RM)**	**PE mark (PM)**	**RM - PM**
1710201	B	2	2	0
1710202	Y	4	5	-1
1710203	U	0	0	0
1710204	H	6	6	0
1710205	A	2	4	-2
1710206	C	3	2	1
1710207	J	3	2	1
1710208	G	4	4	0
1710209	T	4	3	1
1710210	V	0	0	0
1710211	S	5	6	-1
1710212	P	1	0	1
1710213	W	0	0	0
1710214	K	0	0	0
1710215	R	2	0	2
1710216	M	2	2	0
1710217	Q	5	4	1
1710218	O	5	4	1
1710219	N	2	2	0
1710220	L	4	2	2
1710221	F	1	1	0
1710222	X	0	0	0
1710223	D	5	6	-1
1710224	I	0	0	0
1710225	Z	2	2	0
1710226	E	2	0	2

Reliability indicators (see section 4.3.4)



Question	number of scripts	Maximum mark	P_{Agr1}	Mean (RM-PM)	Mean absolute difference
Q1 Polymers	26	6	0.84	0.27	0.65

Kendall's tau correlation

Question 1 Polymers		Examiner mark	Researcher mark
Examiner mark	correlation coefficient Significance	1	0.817 0.00000035
Researcher mark	correlation coefficient Significance	0.817 0.00000035	1

N=26

Shaded boxes show those correlations that are significant at the 0.01 level (2-tailed).

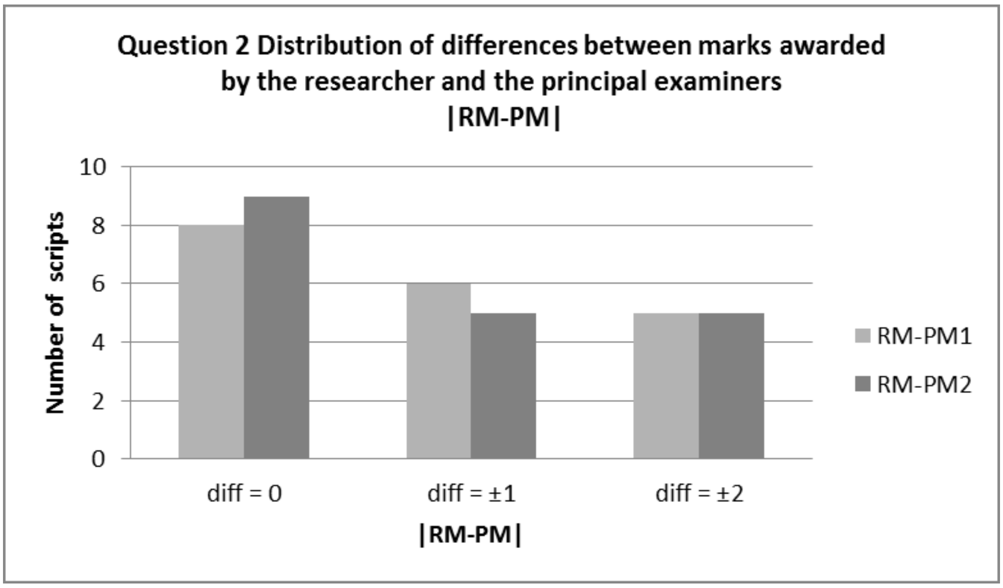
Validation : Question 2 Nuclear power station – A181/2 Q6

Raw data from research

Question 2 Nuclear power station		mark awarded for question using mark scheme derived from theoretical framework				differences between marks awarded		
candidate code	**script code**	**researcher mark (RM)**	**PE1 mark (PM1)**	**PE2 mark (PM2)**	**standardised mark (SM)**	**RM-PM1**	**RM-PM2**	**RM-SM**
1810201	B	2	4	3	3	-2	-1	1
1810202	D	2	1	2	1	1	0	1
1810203	F	2	3	2	2	-1	0	0
1810204	H	2	2	2	2	0	0	0
1810205	A	2	4	3	3	-2	-1	1
1810206	C	3	5	3	3	-2	0	0
1810207	J	2	1	2	1	1	0	1
1810208	E	4	4	4	4	0	0	0
1810209	T	4	4	2	3	0	2	1
1810210	G	3	2	1	2	1	2	1
1810211	S	3	1	1	3	2	2	0
1810212	P	3	3	3	3	0	0	0
1810213	I	3	3	2	2	0	1	1
1810214	K	5	5	5	5	0	0	0
1810215	R	5	5	4	5	0	1	0
1810216	M	5	4	3	4	1	2	1
1810217	Q	5	3	3	4	2	2	1
1810218	O	6	5	6	6	0	-1	0
1810219	N	4	4	3	3	0	1	1
1810220*	L	3	2	2	2	0	0	1

*On inspection, 6 marks should not have been awarded on the original mark scheme. This data was not included in the statistical analysis.

Reliability indicators



Examiner	**number of scripts**	**Maximum mark**	**P_{Agr1}**	**Mean (RM-PM)**	**Mean absolute difference**
PM1	19	6	0.74	0.05	0.79
PM2	19	6	0.74	0.53	0.84

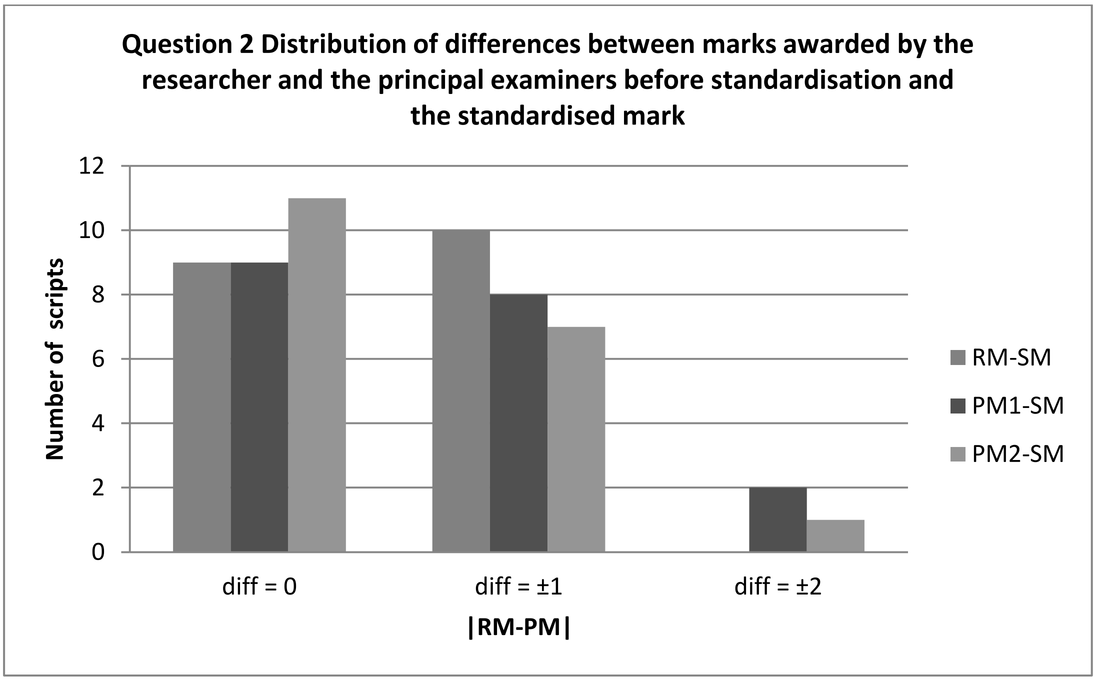
Kendall's tau correlations

Question 2 Nuclear power *N=19*		**Researcher mark**	**PM1**
Researcher mark	correlation coefficient Significance	1	0.506* 0.010
Principal Examiner 1 mark	correlation coefficient Significance	0.506* 0.010	1
		Researcher mark	**PM2**
Researcher mark	correlation coefficient Significance	1	0.489** 0.013
Principal Examiner 2 mark	correlation coefficient Significance	0.489** 0.013	1

* correlation is significant at the 0.01 level (2-tailed).

** correlation is significant at the 0.05 level (2-tailed).

Effect of standardisation



Marker	Number of scripts	Maximum mark	P_{Agr1}	Mean difference between marks awarded before standardisation and the agreed standardised mark	Mean absolute difference between marks awarded before standardisation and the agreed standardised mark
RM	19	6	1.00	0.26	0.58
PM1	19	6	0.89	0.21	0.63
PM2	19	6	0.95	0.47	-0.26

Kendall's tau correlations

Question 2 Nuclear power *N=19*		**SM**	**RM**
SM	correlation coefficient Significance	1	0.758 0.0001
RM	correlation coefficient Significance	0.758 0.0001	1
		SM	**PM1**
Researcher mark	correlation coefficient Significance	1	0.695 0.0003
PM1	correlation coefficient Significance	0.695 0.0003	1
		SM	**PM2**
Researcher mark	correlation coefficient Significance	1	0.750 0.0001
PM2	correlation coefficient Significance	0.750 0.0001	1

Shaded boxes show those correlations that are significant at the 0.01 level (2-tailed).

Appendix 5 Questionnaire

This Appendix includes the letter sent to examiners alongside the background information and questionnaire. For analysis see chapter 6.

UNIVERSITY of York

Science Education Group
Alcuin D Block
Heslington, York YO10 5DD

Telephone: +44(0) 1904 324701
Facsimile: +44(0) 1904 322605
mary.whitehouse@york.ac.uk

20th May 2013

Dear Colleague

GCSE Sciences

I apologise for writing to you at what is, I know, the beginning of a very busy time of year.

I hope you will be able to help me with the research for my MA in Education. My aim has been to find out if it is possible produce a useful framework of descriptors that could be used as the starting point for writing level of response mark schemes for 6-mark questions that asked students to provide arguments or explanations

I began this MA at the time we were first developing the 6 mark questions for the 2009 GCSE Science specifications.

At that same time I was reading the work of some researchers who were developing resources to support teachers in improving students' ability to give a good scientific explanation and that of some other researchers who were developing systems to evaluate students' ability to carry through an argument. As arguing and explaining were both skills that we were expecting of students in some of the questions we were devising, I wanted to find out if we could draw on their ideas in developing mark schemes for some of the 6 mark questions – and also perhaps in writing the questions.

Once the first live examinations had taken place in January 2012 OCR I was able to try out my ideas on some students' work. I would now like to try out my ideas on you!

Attached to this message are two documents:

- a description of my research
- a questionnaire

I hope that you will be willing to read the description of my work and answer the questions in the questionnaire document.

All responses to the questions will be reported anonymously and it will not be possible to identify individuals in the report.

I hope to hear from you soon, ideally within the next month.

Warm regards

Mary
Mary Whitehouse

Encl: *Developing a framework for Level of Response mark schemes for GCSE Sciences*
Questionnaire about the framework for Level of Response mark schemes for GCSE Sciences

Developing a framework for level of response mark schemes for GCSE Sciences

Aim of the research

I have three research questions:

1. Can a theoretically-based model of argumentation be used as the basis for developing a framework to evaluate extended written answers to questions in GCSE Science examinations that require an argument or explanation?
2. Would mark schemes based on such a framework yield comparable outcomes compared with the marks awarded using the conventional mark schemes used by the awarding body?
3. Would examiners find such a framework useful in establishing consistency of demand across different papers within the suite and year on year?

The aim was to produce a framework of descriptors that could be used as the starting point for writing level of response mark schemes for 6-mark questions that asked students to provide arguments or explanations.

This research did not begin from the premise that there was anything wrong with the mark schemes that were being developed for the specimen papers and the first live papers. However a generic framework might help to produce a common approach across papers within a suite, and between examination sessions. In addition, if it could be shared with teachers (and students) it would help them to understand what a good argument or explanation looks like.

Developing the framework

The starting point for developing the framework was the work of Sibel Erduran and her colleagues at Kings College, London (2004). They had researched the use of argumentation during science lessons in secondary schools and had developed a tool to measure the quality of the arguments they observed. This tool was based on the work of Toulmin who had identified the components of an argument (2003). Toulmin describes a sophisticated argument as one that will not only provide evidence (data) to support a point of view (claim), it will also show how that evidence supports the point of view (warrants and backing) and will anticipate an opponent's counterarguments and be able to rebut them.

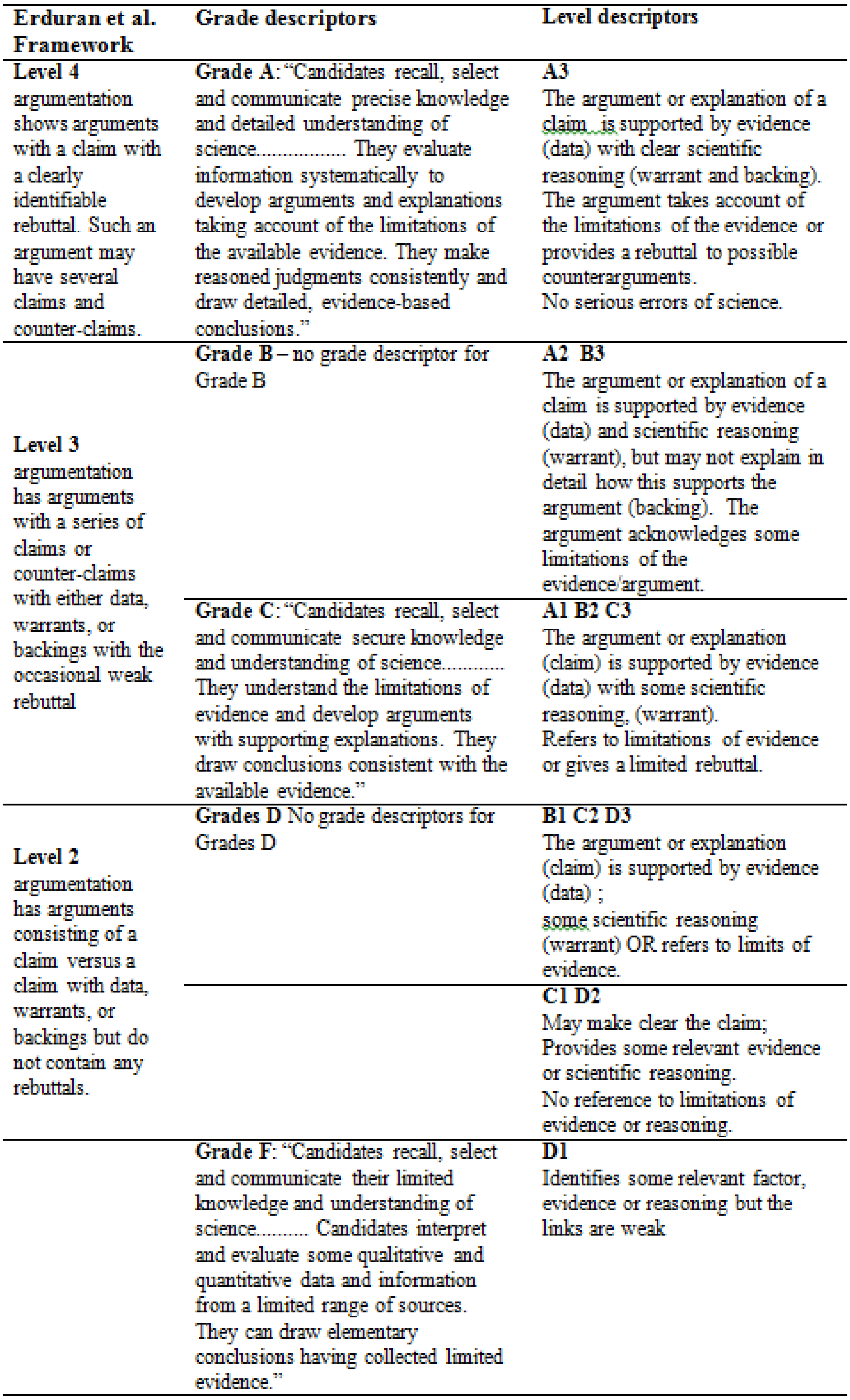
Although we do not talk about argumentation specifically in GCSE Science, the framework used by Erduran to describe different qualities of argument has a hierarchy with similar differentiation between levels to the aspects of Ofqual's grade descriptors that describe the quality of students' explanations (2009). Table 1 shows how the two map against each other.

Erduran Framework	**Grade descriptors**	**Comments**
(Erduran et al., 2004)	**(Ofqual, 2009)**	
Level 5 argumentation displays an extended argument with more than one rebuttal		
Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.	**Grade A**: "Candidates recall, select and communicate precise knowledge and detailed understanding of science................ They evaluate information systematically to develop arguments and explanations taking account of the limitations of the available evidence. They make reasoned judgments consistently and draw detailed, evidence-based conclusions."	"take account of the limitations of the available evidence" considered as equivalent to giving an "identifiable rebuttal" "reasoned judgments consistently and draw detailed, evidence-based conclusions" suggests use of data with scientific 'warrants or backing'
Level 3 argumentation has arguments with a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal	**Grade B** – no grade descriptor for Grade B	
	Grade C: "Candidates recall, select and communicate secure knowledge and understanding of science........... They understand the limitations of evidence and develop arguments with supporting explanations. They draw conclusions consistent with the available evidence."	"and develop arguments with supporting explanations." at this level is equivalent to "arguments with a series of claims ..... with either data, warrants, or backings"
Level 2 argumentation has arguments consisting of a claim versus a claim with data, warrants, or backings but do not contain any rebuttals.	**Grades D and E** No grade descriptors for Grades D and E	
	Grade F: "Candidates recall, select and communicate their limited knowledge and understanding of science.......... Candidates interpret and evaluate some qualitative and quantitative data and information from a limited range of sources. They can draw elementary conclusions having collected limited evidence."	
Level 1 argumentation consists of arguments that are a simple claim versus a counter-claim or a claim versus a claim.		Level 1 is below the description for Grade F.

Table 1 Mapping Erduran framework to Ofqual grade descriptors

Examiners and revisers do not often refer directly to the grade descriptors when setting and revising questions, although they are used when determining grade boundaries at the Award stage. When a question is targeted at grades up to grade A, it might be expected that in a Level of Response (LOR) mark scheme the Level 3 description would match elements of the Ofqual Grade A descriptor. It would then be possible to say that candidates who scored 5 or 6 on the question were producing Grade A quality work. Similarly, for a question targeted at Grade C, the Grade C descriptor would be used as the basis for writing the Level 3 descriptor.

These ideas were used to write generic LOR descriptors that could be used as a framework for writing specific LOR mark schemes for GCSE Science questions. This stage in the development is shown in table 2. **A3**, **A2** and **A1** are the three levels of response that would be expected for a question targeted at Grade A. Similarly **B3**, **B2** and **B1** are the three descriptors for a question targeted at Grade B and so on.



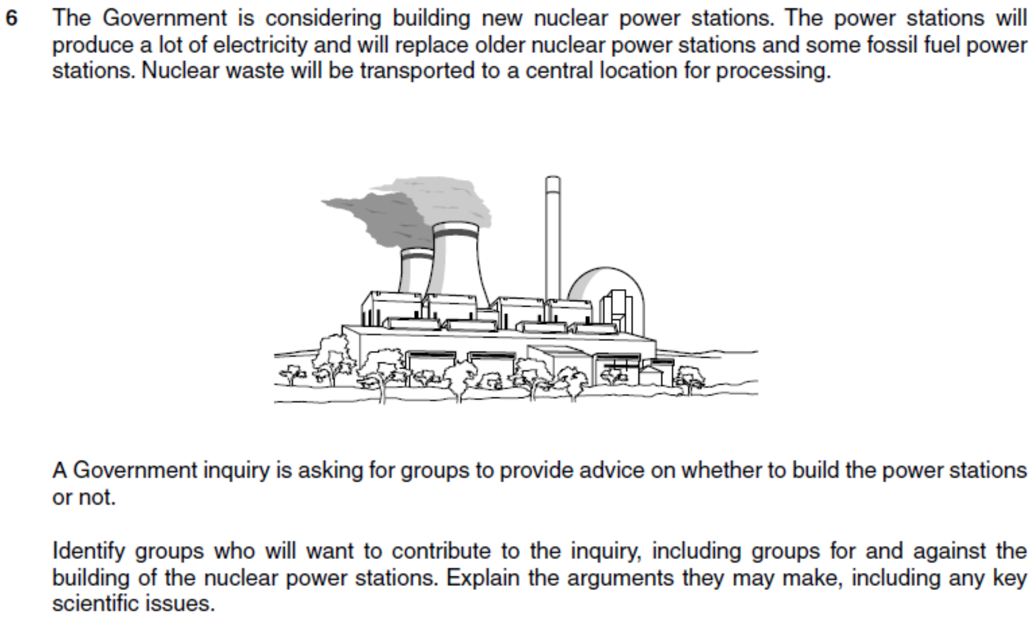
Erduran et al. Framework	**Grade descriptors**	**Level descriptors**
Level 4 argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counter-claims.	**Grade A**: "Candidates recall, select and communicate precise knowledge and detailed understanding of science................. They evaluate information systematically to develop arguments and explanations taking account of the limitations of the available evidence. They make reasoned judgments consistently and draw detailed, evidence-based conclusions."	**A3** The argument or explanation of a claim is supported by evidence (data) with clear scientific reasoning (warrant and backing). The argument takes account of the limitations of the evidence or provides a rebuttal to possible counterarguments. No serious errors of science.
Level 3 argumentation has arguments with a series of claims or counter-claims with either data, warrants, or backings with the occasional weak rebuttal	**Grade B** – no grade descriptor for Grade B	**A2 B3** The argument or explanation of a claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence/argument.
	Grade C: "Candidates recall, select and communicate secure knowledge and understanding of science............ They understand the limitations of evidence and develop arguments with supporting explanations. They draw conclusions consistent with the available evidence."	**A1 B2 C3** The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence or gives a limited rebuttal.
Level 2 argumentation has arguments consisting of a claim versus a claim with data, warrants, or backings but do not contain any rebuttals.	**Grades D** No grade descriptors for Grades D	**B1 C2 D3** The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning (warrant) OR refers to limits of evidence.
		C1 D2 May make clear the claim; Provides some relevant evidence or scientific reasoning. No reference to limitations of evidence or reasoning.
	Grade F: "Candidates recall, select and communicate their limited knowledge and understanding of science.......... Candidates interpret and evaluate some qualitative and quantitative data and information from a limited range of sources. They can draw elementary conclusions having collected limited evidence."	**D1** Identifies some relevant factor, evidence or reasoning but the links are weak

Table 2 Level of response descriptors developed from Erduran framework and Grade descriptors

Applying the framework

The generic framework of level descriptors was then used to write mark schemes for some of the questions set in the January 2012 session. In practice the level of demand for a question is also determined by the demand of the science expected in the response, this is indicated in the additional guidance in the mark scheme.

Figure 1 shows question 6 on the higher tier GCSE Physics paper A181/02 in January 2012. This question was targeted up to Grade B.



6 The Government is considering building new nuclear power stations. The power stations will produce a lot of electricity and will replace older nuclear power stations and some fossil fuel power stations. Nuclear waste will be transported to a central location for processing.

A Government inquiry is asking for groups to provide advice on whether to build the power stations or not.

Identify groups who will want to contribute to the inquiry, including groups for and against the building of the nuclear power stations. Explain the arguments they may make, including any key scientific issues.

Figure 1 Question 6 GCSE Physics paper A181/02 January 2012 (OCR, 2012)

Table 3 shows how the generic level descriptors were interpreted to create a level of response mark scheme for this particular question. The QWC descriptions agreed in January 2012 were included. The guidance in the right hand column identified examples of the science content that is expected for an answer targeted at this level.

Framework descriptors	**Mark scheme Level of Response descriptors**	**Additional Guidance**
B3 The argument or explanation of the claim is supported by evidence (data) and scientific reasoning (warrant), but may not explain in detail how this supports the argument (backing). The argument acknowledges some limitations of the evidence or argument (weak rebuttal).	**Level 3** Answer identifies group **for** and puts forward at least one piece of evidence for with, scientific explanation. Answer identifies group **against** and puts forward at least one piece of evidence against, with scientific explanation. Quality of written communication does not impede communication of the science at this level.	This question is targeted at grades up to B. **Answers at Level 3 must include reference to radioactive materials / ionising radiation**. (As candidates are required to put both sides of the argument, they are not expected to include any explicit rebuttals.) **Groups and arguments for nuclear power station** • environmental groups – reduces CO_2 emissions of power production – so reducing greenhouse gases; reduces particulate/acid rain gases – so reducing environmental damage • local people near old coal stations – less emissions from NPS so cleaner air; nuclear fuel much less bulky, so fewer lorries/rail trucks in and out • workers near PS – provides work during demolition / construction of PS **Groups and arguments against power station** • People living near NPS sites / People near nuclear waste disposal – concerns about ionising radiation during use / risk of accidents – radiation can cause cancer • environmental group – disposal of nuclear waste is a problem: ionising, long lasting
B2 The argument or explanation (claim) is supported by evidence (data) with some scientific reasoning, (warrant). Refers to limitations of evidence.	**Level 2** Answer identifies groups for and against NPS; uses evidence for and against with some scientific reasoning for at least one argument. Quality of written communication partly impedes communication of the science at this level.	
B1 The argument or explanation (claim) is supported by evidence (data) ; some scientific reasoning, (warrant) / refers to limits of evidence.	**Level 1** Puts forward evidence for and against but may not link to groups. Quality of written communication impedes communication of the science at this level.	
	Level 0 Insufficient or relevant science. Answer not worthy of credit.	

Table 3 Generic marking scheme level descriptors and specific descriptors for Q6 A181/02

This mark scheme was used to mark 20 scripts from candidates who took the examination in January 2012. The scheme was also tried out by two of the Principal Examiners for Physics and then we had a 'mini standardisation meeting' where we came to an agreement on the marks we would award using this scheme.

The marks awarded using this scheme were compared with the marks awarded to the same candidates by the original mark scheme in January 2012. For 14 of the 20 scripts the marks awarded were the same or within ±1. Figure 2 shows the distribution of differences between the marks awarded by the original examiner and the mark awarded using the standardised 'research' mark scheme. (On inspection it would seem that the scripts that were awarded +4 and +3 marks by the original examiner had been very generously marked against the requirements of the original mark scheme.)

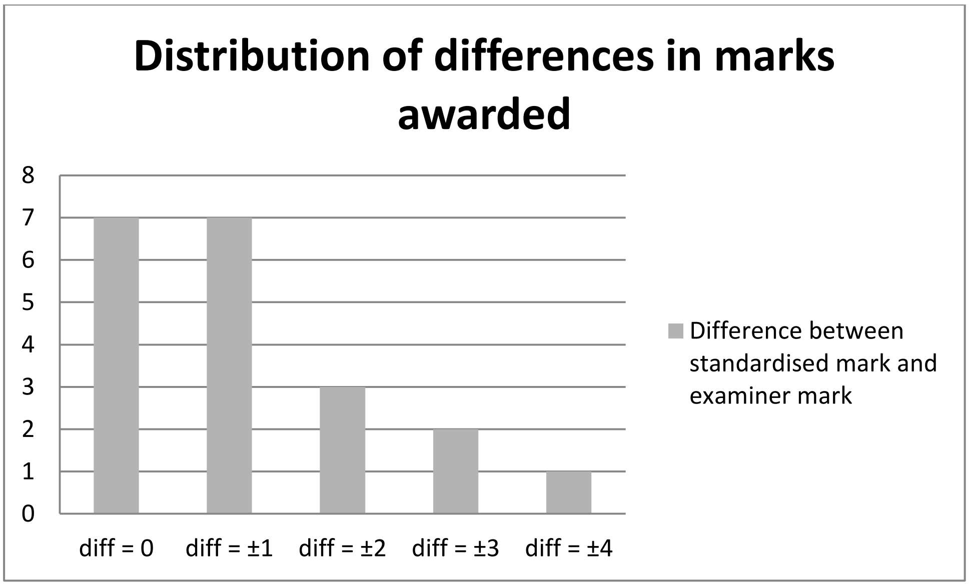


Figure 2: Comparison of marks between those awarded using the original mark scheme with those awarded using the researcher mark scheme after standardisation for Q6 A181/02.

A similar pattern has been seen on the other questions for which the framework has been used, with the mark for most scripts being within ±1 of the mark awarded by the original examiner.

References

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933.

OCR. (2012). GCSE Twenty First Century Science Physics A Higher Tier *A181/02 P1 P2 P3 January*. Cambridge: OCR.

Ofqual. (2009). *GCSE subject criteria for science*. Coventry: Ofqual

Toulmin, S. (2003). *The Uses of Argument* (second ed.). Cambridge: Cambridge University Press.

UNIVERSITY OF YORK – DEPARTMENT OF EDUCATION

RESEARCHER – Mary Whitehouse

DISSERTATION TITLE - DEVELOPING A FRAMEWORK FOR LEVEL OF RESPONSE MARK SCHEMES FOR GCSE SCIENCES

RESEARCH CONSENT FORM

I understand that this research study will investigate whether it is possible to develop a framework of descriptors that could be used as the starting point for writing level of response mark schemes for 6-mark questions that asked students to provide arguments or explanations.

I understand that examiners' involvement in the study is voluntary and therefore can be withdrawn at any time.

I understand that the data and recordings gathered from the activities will be kept confidentially and anonymously and that no unauthorised persons will have access to the data. I understand that no names of examiners will be included in the dissertation.

DECLARATION OF CONSENT

I have been informed about the aims and procedures involved in this research project and by returning the questionnaire I consent to the terms of research, detailed above.

____________________	________________________	_______________
Name	Signature	Date
Mary Whitehouse	*Mary Whitehouse*	20th May 2013
Name of researcher	Signature	Date

Many thanks for your help.

Mary Whitehouse

University of York

Questionnaire for examiners about the framework for Level of Response mark schemes for GCSE Sciences

About you

Please put a tick in the appropriate boxes.

1 Which GCSE Science suites do you work with?

Suite A Twenty First Century Science.................. □

Suite B Gateway Science.................................... □

2 What is your current role in GCSE Science? (Tick all that apply)

Chair / Deputy Chair of Examiners .... □

Chief Examiner.. □

Principal Examiner (setting)................................ □

Principal Examiner (marking) .... □

Reviser.. □

3 How long have you been involved in examining GCSE Science in one or more of the roles listed above?

0 – 5 years □ 6 – 10 years □ More than 10 years □

Please answer the questions on the next page

Level of Response mark schemes

1 Do you think that questions with LOR mark schemes allow examiners to assess skills that are not rewarded using 'conventional' extended answer mark schemes?
YES/NO

Please explain your answer.

....

....

....

....

2 Do you think that the challenges of writing an LOR mark scheme are different from those met when devising a mark scheme for other questions that require an extended written answer?
YES/NO

Please explain your answer.

....

....

....

....

3 Do you think that starting from a common framework based on the grade descriptors would be helpful in devising LOR mark schemes?
YES/NO

Please explain your answer.

....

....

....

....

4 Do you think that the framework devised in this research specifically for questions that ask for explanations or arguments could be useful?
YES/NO

Please explain your answer.

....

....

....

....

Many thanks for your time. Please email your response to: mary.whitehouse@york.ac.uk

Glossary: Technical terms used in this study

Assessors the collective term used in this study when referring to the group of people who are responsible for setting, marking and grading examinations i.e.: **chair of examiners**, **chief examiner, principal examiner, reviser, scrutineer,** and **assistant examiners**

Assistant examiners are responsible for marking the candidates work in accordance with the agreed mark scheme. Where there are a large number of assistant examiners some examiners will also be team leaders, responsible for monitoring the marking of a group of examiners.

Chair of examiners is responsible for maintaining standards across all the specifications within a subject area at an awarding body.

Chief examiner for a specification is responsible for ensuring that all the components of the assessment, both examinations and internal assessment, meet the requirements of the specification and that over a number of examination sessions standards are maintained and all aspects of the specification are assessed.

Cut score is the minimum mark required for a candidate to achieve a particular grade in an examination paper.

Grade descriptors for a qualification describe the characteristics of the performance of a candidate who achieves a particular grade.

Facility of a question is the mean mark awarded for the question as a proportion of the maximum mark for the question. Facility of a question at a grade is calculated using the mean mark for the questions for candidates who achieved that grade on the paper. (Elliott & Johnson, 2007)

Item level data (ILD) provides information about the facility of the question and how it varies for different ability candidates (see section 2.6).

Levels-based mark schemes describe a number of levels of response, each with an associated band of marks**.**

Objective questions – answers to these questions are unambiguous, the mark scheme lists acceptable answers.

Ofqual (Office of Qualifications and Examinations Regulation) regulates qualifications, examinations and assessments in England.

Points-based mark schemes provide a list of acceptable points which must be matched by the candidate's answer.

Principal examiner for a unit of assessment is responsible for setting the paper and mark scheme and standardising the marking of that paper.

Question Paper Evaluation Committee (QPEC) (see section 2.2) the committee that meets to consider drafts of question papers and mark schemes to ensure that they are of high quality and match the specification. This committee is also called the AMEC (Assessment Materials Evaluation Committee).

Revisers provide written comments on early drafts of the paper and mark scheme and attend the QPEC.

Scrutineer checks the final draft of the paper and mark scheme.

Standardisation meeting (see section 2.4) takes place after the exam has been taken and before marking begins. The mark scheme is finalised and additional guidance is added to aid examiners in making decisions when awarding marks. These senior examiners also agree the marks on the scripts that will be used for training, standardisation, and sampling of examiners. This meeting is also called the SSU (Scoris set up meeting), because the SCORIS marking platform is set up at this meeting.

Tariff of a question is the maximum mark that could be awarded for that question.

References

AQA, (Assessment and Qualifications Alliance). (2011a). GCSE specification Science A. Manchester: AQA.

AQA, (Assessment and Qualifications Alliance). (2011b). GCSE specification Science B (Science in context) 4500. Manchester: AQA.

AQA, (Assessment and Qualifications Alliance). (2013). Do I qualify to be an examiner or moderator? Retrieved 25th July 2013, from http://www.aqa.org.uk/about-us/work-with-us/examiners-and-moderators/do-i-qualify-to-be-an-examiner-or-moderator

Baird, J.-A., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). *Research evidence relating to proposals for reform of the GCSE*. Oxford: Oxford University Centre for Educational Assessment.

Baird, J.-A., & Black, P. (2013). The reliability of public examinations. *Research Papers in Education, 28*(1), 1-4.

Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331-348.

Bell, J. F., Bramley, T., Claessen, M. J. A., & Raikes, N. (2007). Quality control of examination marking. *Research matters*(4), 4.

Berland, L. K., & McNeill, K. L. (2012). For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson. *Science Education, 96*(5), 808-813.

Biggs, J. (1998). Assessment and Classroom Learning: a role for summative assessment? *Assessment in Education: Principles, Policy & Practice, 5*(1), 103-110.

Black, B., Sütő, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18*(3), 295-318.

Black, P. (1998). *Testing: friend or foe? Theory and practice of assessment and testing*. London: Falmer Press.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues *Research matters* (Vol. 4). Cambridge: Cambridge Assessment.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh. http://www.cambridgeassessment.org.uk/images/109770-mark-scheme-features-associated-with-different-levels-of-marker-agreement.pdf

CA, (Cambridge Assessment). (2012a). *Unit A141/01: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012b). *Unit A141/02: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012c). *Unit A142/01: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012d). *Unit A142/02: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012e). *Unit A161/01: January 2012: Post Awarding Item Level Data Report*. Cambridge.

CA, (Cambridge Assessment). (2012f). *Unit A161/02: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012g). *Unit A171/01: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012h). *Unit A171/02: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012i). *Unit A181/01: January 2012: Post Awarding Item Level Data Report* Cambridge.

CA, (Cambridge Assessment). (2012j). *Unit A181/02: January 2012: Post Awarding Item Level Data Report* Cambridge.

Chamberlain, S. (2008). Do marking reliability studies have validity? Manchester: AQA Centre for Education Research and Policy.

Chen, Y.-C. (2011). *Examining the integration of talk and writing for student knowledge construction through argumentation*. doctoral PhD dissertation. University of Iowa. Iowa. Retrieved from http://ir.uiowa.edu/etd/1129.

Department for Education and Employment. (2001). *Key Stage 3 National Strategy: Literacy across the curriculum*. London: DfEE.

Department for Education and Science. (1989). *Science in the National Curriculum*. London: HMSO.

Department for Education and Skills. (2002). *Key Stage 3 National Strategy Literacy in science*. London: DfES.

Doyle, W., & Ponder, G. (1977). The practicality ethic in teacher decision-making. *Interchange, 8*(3), 1-12.

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

Duschl, R. A., & Osborne, J. (2002). Supporting and Promoting Argumentation Discourse in Science Education. *Studies in Science Education, 38*(1), 39-72.

Edexcel. (2013a). Examiner. Retrieved 25th July 2013, from http://www.edexcel.com/i-am-a/teacher/aa-recruit/roles/Pages/Examiner.aspx

Edexcel. (2013b). *GCSE Sciences Edexcel GCSE in Sciences 2SC01*. Harlow: Pearson Education.

Elliott, G., & Johnson, N. (2007). Item Level Data: Guidelines for Staff. UK: Cambridge: Cambridge Assessment.

Erduran, S., & Jiménez-Aleixandre, M. P. (2008). Argumentation in science education: An overview. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research* (pp. 3-27). Dordrecht, The Netherlands: Springer.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications.

Fullan, M. (2007). *The new meaning of educational change*. Abingdon: Routledge.

Greatorex, J. (2003). Developing and applying level descriptors. *Westminster Studies in Education, 26*(2), 125-133.

Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). “Doing the lesson” or “doing science”: Argument in high school genetics. *Science Education, 84*(6), 757-792.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving Score Differences in the Rating of Writing Samples: Does Discussion Improve the Accuracy of Scores? *Language Assessment Quarterly, 2*(2), 117-146.

Joint Council for Qualifications. (2013). Examiners and marking. Retrieved 25th July 2013, from http://www.jcq.org.uk/examination-system/the-role-of-an-examiner

Kelly, G. J., Druker, S., & Chen, C. (1998). Students’ reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education, 20*(7), 849-871.

Khishfe, R. (2013). Explicit Nature of Science and Argumentation Instruction in the Context of Socioscientific Issues: An effect on student learning and transfer. *International Journal of Science Education*, 1-43.

Kind, P. M., Kind, V., Hofstein, A., & Wilson, J. (2011). Peer argumentation in the school science laboratory — exploring effects of task features. *International Journal of Science Education, 33*(18), 2527-2558.

Knight, A. M., & Grymonpre, K. (2013). Assessing students' arguments: How strong are their justifications? *Science Scope, 36*(9), 51-59.

Krajcik, J. S., & McNeill, K. L. (2009). *Designing instructional materials to support students’ in writing scientific explanations: Using evidence and reasoning across the middle school years*. Paper presented at the National Association for Research in Science Teaching, Garden Grove, California.

Massey, A. J., & Raikes, N. (2006). *Item-level examiner agreement*. Paper presented at the Annual Conference of the British Educational Research Association, University of Warwick. http://www.cambridgeassessment.org.uk/Images/111065-item-level-examiner-agreement.pdf

McNeill, K. L., & Krajcik, J. S. (2011). *Supporting grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing*. Upper Saddle River: Pearson.

Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. Report for the National Assessment Agency: AQA Centre for Education Research and Policy.

Millar, R. (2006). Twenty First Century Science: Insights from the design and implementation of a scientific literacy approach in school science. *International Journal of Science Education, 28*(13), 1499-1521.

Millar, R. (2013). Improving science education: Why assessment matters. In D. Gunstone, D. Corrigan & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy*. Dordrecht: Springer.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future* (S. o. E. King’s College London Ed.). London: Kings College London.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington DC: National Academies Press.

Newton, P. (2009). The reliability of results from national curriculum testing in England. *Educational Research, 51*(2), 181-212.

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553-576.

OCR, (Oxford, Cambridge and RSA Examinations). (2005). *Twenty First Century Science Suite: GCSE Science A J630*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2011). *Twenty First Century Science Suite: GCSE Science A J241*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012a). *Biology A J243 OCR Report to Centres January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012b). *Chemistry A J244 OCR Report to Centres January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012c). *Gateway Science GCSE Science B*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012d). *GCSE Twenty First Century Science Biology A Foundation Tier A161/01 January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012e). *GCSE Twenty First Century Science Biology A Foundation tier A161/01 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012f). *GCSE Twenty First Century Science biology A Higher tier A161/02 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012g). *GCSE Twenty First Century Science chemistry A Foundation tier A171/01 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012h). *GCSE Twenty First Century Science Chemistry A Foundation Tier A171/01 January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012i). *GCSE Twenty First Century Science Chemistry A Higher Tier A171/02 January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012j). *GCSE Twenty First Century Science Chemistry A higher tier A171/02 January 2012 mark scheme* Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012k). *GCSE Twenty First Century Science physics A foundation tier A181/01 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012l). *GCSE Twenty First Century Science physics A Higher tier A181/02 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012m). *GCSE Twenty First Century Science science A Foundation tier A141/01 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012n). *GCSE Twenty First Century Science science A Foundation tier A142/01 January 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012o). *GCSE Twenty First Century Science science A Higher tier A141/02 Janury 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012p). *GCSE Twenty First Century Science science A Higher tier A142/02 Janury 2012 mark scheme*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012q). *GCSE Twenty First Century Science Physics A Foundation Tier A181/01 January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012r). *GCSE Twenty First Century Science Physics A Higher Tier A181/02 January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2012s). *Physics A J245 OCR Report to Centres January 2012*. Cambridge: OCR.

OCR, (Oxford, Cambridge and RSA Examinations). (2013a). Active results. from http://www.ocr.org.uk/ocr-for/teachers/active-results/

OCR, (Oxford, Cambridge and RSA Examinations). (2013b). Become an assessor. Retrieved 25th July 2013, from http://www.ocr.org.uk/ocr-for/assessors/become-an-assessor/

Ofqual, (Office of Qualifications and Examinations Regulation). (2008). *Reliability programme*. Coventry.

Ofqual, (Office of Qualifications and Examinations Regulation). (2009). *GCSE subject criteria for science*. Coventry: Ofqual.

Ofqual, (Office of Qualifications and Examinations Regulation). (2011). *GCSE, GCE, Principal learning and project code of practice*. London: HMSO Retrieved from www.ofqual.gov.uk/publications.

Ofqual, (Office of Qualifications and Examinations Regulation). (2012). *GCSE Science chair of examiners report template*. Coventry: Ofqual Retrieved from http://www2.ofqual.gov.uk/downloads/category/131-guidance.

Ofqual, (Office of Qualifications and Examinations Regulation). (2013a). About marking and grading. Retrieved 22nd July 2013, 2013, from http://ofqual.gov.uk/help-and-advice/about-marking-and-grading/

Ofqual, (Office of Qualifications and Examinations Regulation). (2013b). *Reforms to GCSEs in England 2015*. Coventry: Retrieved from http://comment.ofqual.gov.uk/gcse-reform-june-2013/.

Ofqual, (Office of Qualifications and Examinations Regulation). (2013c). *Review of Quality of Marking in Exams in A levels, GCSEs and Other Academic Qualifications Interim report* (Crown Ed.). Coventry: Ofqual.

Ofqual, (Office of Qualifications and Examinations Regulation). (2013d). *Summer 2013 Data exchange procedures*. Coventry: Ofqual.

Osborne, J. (2011). Science teaching methods: A rationale for practices. *School Science Review, 93*(343), 93-103.

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

Osborne, J., Erduran, S., & Simon, S. (2004b). *Ideas, evidence and argument in science. In-service training pack, resource pack and video*. London: Nuffield Foundation.

Osborne, J., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education, 95*(4), 627-638.

Osborne, J., & Patterson, A. (2012). Authors' response to “For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson” by Berland and McNeill. *Science Education, 96*(5), 814-817.

QCA, (Qualifications and Curriculum Authority). (1999). *The National Curriculum for England : Science*. London: DfEE, QCA.

Robson, C. (2011). *Real world research: A resource for users of social research methods in applied settings*. Chichester: Wiley.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Simon, S., & Richardson, K. (2009). Argumentation in school science: Breaking the tradition of authoritative exposition through a pedagogy that promotes discussion and reasoning. *Argumentation, 23*(4), 469-493.

Spear, M. (1997). The influence of contrast effects upon teachers’ marks. *Educational Research, 39*(2), 229-233.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161-179.

Sütő, I., & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23*(4), 477-497.

Sütő, I., Nádas, R., & Bell, J. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21-51.

Tiberghien, A. (2008). Foreword. In S. Erduran & M. P. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research*. Dordrecht, The Netherlands: Springer.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A Review on Marking Reliability Research*. Slough: NFER.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Toulmin, S. (2003). *The Uses of Argument* (2nd ed.). Cambridge: Cambridge University Press.

Twenty First Century Science. (2006). *Twenty First Century Science. GCSE Science Higher and Foundation*. Oxford: Oxford University Press.

Walton, D. N. (1996). *Argument structure: A pragmatic theory*. Toronto: University of Toronto Press.

Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. Buckingham: Open University Press

Wenger, E. (2000). Communities of Practice and Social Learning Systems. *Organization, 7*(2), 225-246.