Qualifications and Curriculum Authority

Inter-subject comparability studies

February 2008

QCA/08/3568

Contents

Executive summary

Four investigations, or studies, were carried out to address issues regularly raised on standards across subjects at A level. The scope of two of these studies also extended to GCSE level. Subject experts, with a background in assessment, were employed as reviewers to analyse assessment materials and candidate work across two or more cognate subjects, to draw comparisons and highlight differences in demand.

The four studies focused on the following areas:

Study 1a: Comparisons at GCSE, AS and A level using selected specifications across geography and history

Study 1b: Comparisons at GCSE, AS and A level using selected specifications across biology, chemistry and physics and, at GCSE, science (double award)

Study 2a: Comparisons at A level using selected specifications across biology, psychology and sociology

Study 2b: Comparisons at A level using selected specifications across English literature, history and media studies

Each study involved an evaluation of the demands implied by the syllabus materials within each subject at each level, and a comparison of candidates' work across each subject. The outcomes of each study are outlined in this report.

For study 1a, overall, variations in the schemes of assessment were considered a more important influence on variations in demand than the intrinsic characteristics of the two subjects. The GCSE history assessment was judged to be somewhat more demanding than geography, especially for foundation candidates. At both GCSE and AS, history was judged to be more demanding than geography. However, overall the two subjects were considered to be in line with each other in terms of demand for the A level.

Overall, reviewers found that at A level the specifications had very different assessment characteristics, with an emphasis on different assessment objectives, but that overall the specifications placed similar demands on candidates.

In every case except geography at AS, reviewers rated both subjects as slightly too demanding, with GCSE history being seen as significantly demanding.

During the review of candidates' work, reviewers found that geography candidates at both GCSE foundation and higher tier showed much less evidence of attainment than corresponding candidates in history. This was also true to a lesser extent at AS. Only in the work on the A2 papers was performance judged to be comparable. Reviewers noted that these differences arose from significant variation in the way the two subjects were assessed. In geography, there was a preponderance of short-answer questions, which focused on very specific items of knowledge. For any question where candidates did not possess that knowledge, they had very little opportunity to show their further knowledge, understanding and skills.

Study 1b concluded that chemistry was the most demanding of the selected subjects at all three qualification levels and that progression between qualification levels was not comparable across the subjects. Several other differences were perceived as having an effect on levels of demand: the options within subjects, the range of questions and question types, the way in which marks were allocated, the approach to the assessment of quality of written communication and, at A level, the approach to the assessment of the ability to synthesise knowledge, understanding and skills (assessment objective 4). At GCSE, there were also issues raised about the way in which tiering worked.

The review of performance standards revealed no substantial or consistent differences in standards of performance between any subjects at any level. This was particularly true at GCSE, where the differences were very small. At A level the picture was slightly less consistent. At AS, the differences were again very small, with biology and chemistry almost indistinguishable and performance in physics judged to be marginally less impressive. At the lower grades in A2, the picture was again one of great consistency, with all three subjects almost perfectly aligned. At the higher grades, however, some divergence was found, with biology candidates performing slightly less well in order to achieve their grades than the chemistry candidates, with the physics candidates in between.

Overall in study 2a, reviewers judged that biology demanded the greatest breadth of detailed knowledge, requiring students to demonstrate high levels of recall. It did not, however, make the same demands in terms of evaluation/interpretation as sociology or psychology.

Sociology was judged to be, potentially, very demanding, because of the requirement to contextualise judgements in appropriate theory. There was concern, however, that non-contextualised, common sense responses could receive too much credit and that the most popular route through the AS would give candidates too much opportunity to write

uncritically, from their own experience**.** The lack of prescription and apparent leniency of the mark schemes added to concern.

Psychology did not give candidates credit for anecdotal knowledge. Instead, the mark schemes made clear demands for candidates to use correct technical terminology and couch their answers in appropriate psychological theory. Psychology was judged to be technically demanding.

Psychology and biology question papers made use of complex concepts. Candidates could neither infer nor guess answers and this rigour was maintained by demanding mark schemes. This was not clearly the case in sociology.

The review of performance standards at AS found all three subjects very well aligned across the grade range. At A2 there was some evidence that the performance of sociology candidates was not as impressive as that of candidates in either psychology or biology, although the nature of the work used in the study makes it hard to gauge how much weight to place on this finding.

Overall the analysis suggested that standards in biology and psychology were very well aligned across the grade range in both the AS and A2 examinations. Given that the initial impetus for the work was the suggestion that students were turning away from science to psychology because it was perceived to be the soft option, the study suggests that this perception has little basis in fact, at least in terms of the demand of the examinations and the grading standards set.

Study 2b concluded that overall there was no clear evidence of significant differences in demand between the three subjects.

Where options were offered in the examination papers for history and media studies, there appeared to be only minor inconsistencies in demand. This was less clearly the case for English literature, where uneven demand was a concern.

The review of performance standards showed there was very little difference between the subjects at the grade A boundary at AS, while at the grade E boundary the performance of candidates in media studies was considered to be slightly less secure than that of candidates in English, with history candidates in between. At A2, at both boundaries, the media studies candidates were considered to be less impressive than the English candidates, with the history candidates in between. There was concern that media studies allowed candidates to use prepared materials in some of the assessment activities and some evidence of this was seen in the script review.

However, it must be recognized that the media studies work seen by reviewers did not include candidates' coursework, a substantial part of the work produced by candidates for their A level. This work displayed a significantly different set of skills from those required in English and history.

1. Introduction

From time to time, concerns are raised about whether the standards required to achieve success in GCSE and GCE A levels are the same across different subjects. The basis of such concerns varies. At their simplest, they derive from the numbers of candidates succeeding in the different subjects. For example, the argument was put forward in August 2003 that the 20 per cent failure rate in AS mathematics showed that it was a harder subject than English, which had a failure rate of only 5 per cent. (Interestingly, such arguments did not mention that 27 per cent of candidates got an A in AS mathematics, while only 16 per cent did so in English. Nor did they use the A level figures: mathematics, 39 per cent A, 95 per cent pass; English, 20 per cent A, 98 per cent pass – all of which seem to tell a rather different story.)

There are more sophisticated approaches to the use of numbers, but these still depend on a purely statistical analysis, using measures of either prior or concurrent attainment to make the comparison. Sometimes, one hears a comment that implies something rather more qualitative (students find subject X harder than subject Y) but the actual origin of these claims is seldom clear. It seems likely these are either anecdotal or, more commonly, small-scale versions of the statistical approach. That is, ‘students at my school tend to get better results in one subject than in another, so one subject is easier than the other’.

One of the occasions when the issue became a focus of concern was on the publication of the Dearing's *Review of qualifications for 16- to 19-year-olds* in 1996. This made use of analyses carried out by a team of researchers led by Professor Carol Fitz-Gibbon at the University of Newcastle (and, latterly, Durham). They used GCSE data to determine the typical relationship between GCSE performance and performance at A level. They then used the profile of GCSE performance for separate subjects to predict the expected A level results, subject by subject, and compared them with the actual results. Differences between predicted and actual outcomes were presented as ’correction factors’ that would need to be applied to bring subjects into line.

The approach is certainly attractive. Such research is relatively cheap and easy to carry

out, even on a very large scale.[1] It also produces results that provide a comforting sense of accuracy, commonly being reported to two decimal places.

The approach depends on a particular interpretation of examination results, one in which they are all essentially measures of aptitude or general ability. This interpretation is important. It is an inevitable part of what examinations do measure, so that results across a range of subjects do give a good measure of general ability. The interpretation is also of practical importance, as it has the potential to affect subject choices, with learners avoiding subjects that don't reward equivalent levels of aptitude with equivalent results.

However, there are other arguments to consider. The first is that where subject choices are being made, this measure of subject difficulty will be only one factor, and probably not the most important one, in the equation. Personal experience of difficulty will often differ from the statistical model. For example, learners who do well at science are often frustrated by the lack of a right answer that characterises many humanities subjects. Enjoyment of a subject is also likely to be important as is the relevance of a subject for career goals.

It is also important to understand other features of comparisons of subject difficulty based on general ability. The main one is the approach assumes that the relationship between each subject and general ability is the same. Yet this rather denies the need for subjects to have characteristic features or a characteristic balance between knowledge, conceptual understanding, cognitive skills and practical skills.

It is also interesting to note that this method produces slightly different outcomes for identifiable sub-groups (males/females or different types of centre, for example), although there is no obvious reason why that should be so[2]. Outcomes also vary slightly according

---

[1] Shortly after the Dearing Report, the Joint Council for General Qualifications (now the Joint Council for Qualifications) commissioned a similar analysis with results that were broadly in line with the earlier findings. This analysis used data comprising essentially the whole population of 18-year-old A level candidates which meant that there was no possibility of sampling error.

[2] An interesting case study here is GCSE mathematics and English. Here subject pairs analysis – ie using concurrent attainment – suggests that mathematics is somewhat harder than English. Because both subjects are taken by most of the age cohort, the relative difficulties are reflected in the raw grade distributions. Unfortunately, as the grade distributions also show, there is a considerable gender issue: boys' success rates in mathematics and in English are very similar; girls achieve much greater success in English than in mathematics. It is hard, in these circumstances, to make sense of the data for the whole cohort.

to small variations in the precise details of the analysis. However, there is no obvious argument for one method of analysis being the best, making the choice of method essentially arbitrary. Most interestingly, similar analyses suggest that the pattern of outcomes is replicated fairly consistently across the English-speaking developed world, while there are quite striking differences for cultures that have appreciably different attitudes to subjects like mathematics and science.

There is one final problem. The use of the term ‘correction factors’ makes people assume that the source of the differences lies in grading standards, and what is required is that standards in all subjects are adjusted to bring them into line. This serves to make people unhappy with the current system, but it is not clear how the differences can be resolved. Importantly, affected subject communities would find it hard to accept the potentially considerable changes that would be needed to achieve the alignment. And this says nothing of the effect on the cohorts of candidates whose results would be affected.

At the very least, therefore, there are good reasons to suggest that there should be some attempt to take account of all those factors that might be contributing to the measured differences, to ensure their full extent is understood. The problem is that this would be complex and expensive (requiring extensive attitude analysis, for example). Because of these reservations, the Qualifications and Curriculum Authority (QCA) has instead tried to develop a different, more qualitative approach to the topic.

Such an approach is by no means straightforward. It is highly resource-intensive and therefore inevitably expensive. It is also not immediately possible to make comparisons across the full range of subjects. The method depends on finding subject experts who are sufficiently qualified to make comparisons across subjects. This almost inevitably rules out comparing, say, French and chemistry directly. Instead, comparisons have to be made between more closely related subjects, where one can reasonably expect there to be teachers and other experts with sufficient expertise in both.

The approach was piloted looking at A levels in economics and business studies and GCSEs in modern foreign languages. This work suggested that the approach was workable and provided the opportunity to refine the methodology. At about this time, an independent panel of experts reviewing the means of maintaining A level standards recommended that QCA ’conduct qualitative analyses, in two subjects, of a series of examinations and resulting scripts detailing content and cognitive requirements’ (*Maintaining GCE A level standards: the findings of an independent panel of experts,* 2002, p24). Accordingly, the decision was taken to go ahead with the first full study, an investigation into geography and history, looking as GCSE and A level. This was intended

to be the first stage in creating a matrix of overlapping work, allowing for all subjects to be compared, if only by inference. The second stage, covering science subjects, again at both GCSE and A level, started a year later.

As this work was progressing, the question of inter-subject comparability again came under the spotlight in August 2003, when the publication of the A level results was accompanied by questions about the possible impact of perceptions of 'soft' subjects. Two subjects were identified as being of particular concern: psychology and media studies. As a result, it was decided to carry out two further studies covering these subjects. One looked at biology, psychology and sociology; the other looked at English literature, history and media studies.[3] These exercises were different from the original studies in two respects: they covered A level only and they looked at three distinct subject areas. These differences had some impact on the methodology, as did a process of review by an expert group QCA assembled to advise on its work in standards and comparability.

This report describes the outcomes of all four reviews. For this purpose, the review of GCSE and A level geography and history will be referred to as study 1a; that of GCSE and A level science examinations as study 1b; that of A level biology, psychology and sociology as study 2a; and that of A level English literature, history and media studies as study 2b. This report covers all four studies and any key findings emerging from them.

It is important to note that the methodology has evolved throughout the process. Readers should therefore refer to the specific studies for full details of the methodology employed, as this may in cases have differed from the general description provided in this main report.

[3] It should be noted that each of the two later studies provides an overlap with the planned programme of work. Biology, covered by study 2a, also forms part of study 1b, while history is included in both study 1a and study 2b.

2. Methodology

The account of the methodology that follows covers two main areas. Firstly, it looks at where the approach was common across all the studies. Secondly, it highlights some aspects that were particular to one of more of the studies, but that emphasize important features of the work or that have an important bearing on the findings. Separate reports on each study, to be published separately, provide further details of the specific methodology used.

2.1 Scope

There were two main components of this work on examinations:

- an analysis of specification materials and an evaluation of the demands of each subject for each qualification
- a comparison of the work of candidates within each subject.

Within study 1a, there was an attempt to develop a taxonomy of common level descriptors for geography and history. The main focus of this taxonomy was to describe performance across levels of the qualifications framework, and it remained in use in the science study, where the issue of performance across levels remained a significant focus of the work. However, this was not a major consideration in studies 2a and 2b, where the taxonomy was not used.

2.2 Personnel

Consultants were recruited through a combination of advertisement and recommendation, to form a balanced team. The aim was to find participants who had experience of teaching more than one subject, and of teaching at least one subject at A level. Knowledge of the examination system was an advantage, but not essential. Almost inevitably, each participant had a main subject specialism, but was able to compare that subject with others in the study. To avoid bias, the aim was to have the main subjects evenly represented across the teams, although this was not achieved for studies 2a and 2b.

Lead consultants were also appointed. Their role was to assist in the development of the various instruments used in the studies, to advise consultants on subject-specific matters and lead subject-specific discussions at meetings, and to prepare the subject-specific parts of the reports.

2.3 Choice of specifications

In all studies, the review focused on a single specification for each subject. The specifications were selected on the basis of size of candidate entry, the general principle being to select the specification with the highest entry. There were two exceptions to this principle: OCR B (Avery Hill) GCSE geography has the second highest GCSE geography candidature, but was selected in preference to AQA A (with the highest candidature) because of its closer similarity in type of scheme of assessment to Edexcel GCSE history, and in approach to Edexcel A level geography specification B. Similarly, the highest entry specification for A level history is from Edexcel but, as a significant amount of work (including study 1a) had focused on this specification in the past, it was decided to use the specification with the second highest candidature, OCR history (3835/7835), in study 2b. This also allowed some reflection on whether variations between specifications within a subject may have affected the outcomes of the work.[4] Table 1 identifies the specifications used.

Table 1

Study 1a				
	Geography		**History**	
GCSE	OCR Avery Hill (1587)		OCR The Modern World (1607)	
A level	Edexcel (8215/9215)		Edexcel (8264/9264)	
Study 1b				
	Biology	**Chemistry**	**Physics**	**Double award**
GCSE	AQA (3411)	AQA (3421)	AQA (3451)	AQA (3462)
A level	OCR (3881/7881)	AQA (5421/6421)	Edexcel (8540/9540)	N/A

[4] This was an issue in the pilot work on A level business studies and economics and across modern foreign languages at GCSE. In the latter investigation in particular, patterns across languages within an awarding body were quite different.

Study 2a			
	Biology	**Psychology**	**Sociology**
A level	OCR (3881/7881)	AQA (5181/6181)	AQA (5191/6191)
Study 2b			
	English literature	**History**	**Media studies**
A level	AQA (5741/6741)	OCR (3835/7835)	OCR (3860/7860)

2.4 Instruments and materials for the review of examination demand

The methodology was designed to provide a framework with common points of reference to compare the demand of the specifications and their associated examinations across the relevant subjects and, for studies 1a and 1b, to enable comparison across different levels of the qualifications framework. The methodology included the use of the forms described in Table 2.

Table 2

Each member of the review team was to use the following	
Form A	For factual description and analysis of specifications, question papers and mark schemes. One to be completed for each specification
Form B	For identification and evaluation of differences in demand between the relevant specifications
CRAS form	For analysis of question papers and mark schemes in terms of complexity, resources, abstractness and strategy (CRAS). One to be completed for each externally assessed unit of each specification reviewed

In addition, studies 1a and 1b made use of the taxonomy of examination demand. This was redrafted at various stages, and in particular at the end of the review of candidate

performance. The final version of the taxonomy from study 1a will be provided in as an appendix to the detailed report on that study, to be published later.

2.5 Process for review of examination demand

2.5.1 Initial briefing

Prior to the commencement of their work, all reviewers in a given study attended an initial briefing. The purpose of this briefing was to familiarise reviewers with the methodology and the instruments used. From the beginning it was stressed that the full review would include scrutiny of candidates' work and that reviewers would have the opportunity to revisit their judgements in the light of that work. Initially, however, reviewers were asked to focus their attention on comparability between specifications and assessment materials in terms of the demands that these would make on candidates.

Throughout the process it was recognised that the use of common terminology across subjects might conceal differences in the use of those terms in particular contexts; conversely, the use of different terminology to describe similar concepts might conceal areas of actual commonality. For example, while a careful consideration of the assessment objectives for each subject and their relative weightings would constitute the first step in any such exercise, it was quite possible that the subjects in the study employed different interpretations of knowledge-based and skills-based constructs identified in their particular assessment objectives. In addition, it was recognised that, even if common knowledge-based or skills-based constructs could be identified, the particular ways in which the assessment materials addressed the assessment objectives could make direct comparisons difficult. This proved a significant factor in study 1a.

2.5.2 Forms A, B and C

Form A

Reviewers used Form A to provide a factual analysis of the specifications, question papers and mark schemes. A generic Form A, as used across QCA review work, was considered by the lead consultant(s) and slight alterations were made to the questions to focus reviewers' attention more clearly on issues relevant to the particular nature of the task.

For studies 2a and 2b**,** there was a major modification to the way in which Form A was used, as a result of input from the expert group. Previously, each reviewer had completed a detailed factual analysis of the specification and assessment materials, responding to prompts on Form A and logging their responses on the form. For studies 2a and 2b, the

majority of the factual analysis was completed by QCA staff as a desk research exercise, and was provided on Form A. When QCA asked reviewers to check the accuracy of the information and comment on features relevant to the study, this modification was judged to be very successful, as it enabled reviewers to focus their time and attention on making judgements about issues relating to standards and comparability.

Form B

Reviewers used Form B to identify differences in demand between the subjects they were reviewing at individual qualification level. One Form B was completed by each reviewer for each pair of review subjects/levels. Reviewers were asked to complete Form B in the light both of their comments on Form A and their completed CRAS forms (see below).

Reviewers used a five-point numerical scale to make judgements about demand, ranging from 1 (very undemanding), to 3 (about right) to 5 (very demanding) to assess the qualification for each subject reviewed. After making each numerical judgement, each reviewer was asked to give a brief summary of the reasons for that judgement. Reviewers then used these numerical judgements and their explanatory comments to make comparative summaries of the demand in the two qualifications. Each reviewer came to a conclusion about overall demand.

Form C

Studies 1a and 1b used a further form – Form C. This was completed by each reviewer as a summary of all judgements, to allow reviewers to gain an easy overview of their pattern of judgements across levels as well as between subjects. Reviewers transferred the numerical judgements made on Form B for each individual qualification to Form C and then, where necessary, added summative comments. The sections in Form C were identical to the sections in Form B.

For studies 2a and 2b, it was decided not to use Form C, since only one level of the National Qualifications Framework was under consideration. If a summative document had been found to be necessary, this could have been generated by QCA.

2.5.3 CRAS forms

The CRAS forms were used to enable the reviewers to reach judgements about the demand of the question papers, based on the nature of the questions. Reviewers were asked to assess the extent to which the question papers made demands in terms of: the complexity of the processes required to answer a question; the extent to which the resources needed to answer the question were provided on the paper; the level of

abstractness of questions; and the extent to which candidates were required to generate a strategy in their answers.[5] To do this, they used a numerical scale.

Reviewers were given a detailed explanation at the initial briefing about each aspect of the CRAS analysis and there was a general discussion about the ways in which the demands of a particular question could be manipulated by making adjustments to the question in terms of complexity, resources, abstractness or strategy.

2.5.4 Numerical scale

The studies used different scales, to reflect the fact that studies 1a and 1b were looking across all three levels of the National Qualifications Framework covered by GCSE and GCE A level, whereas studies 2a and 2b looked at A level only. Because A level now comprises two separate standards, AS and A2, effectively studies 1a and 1b covered four different standards (GCSE foundation, GCSE higher, AS and A2), while studies 2a and 2b covered AS and A2. For studies 1a and 1b, reviewers used a ten-point scale; for the other studies they used a six-point scale. These scales were judged to provide the flexibility necessary for the task, without being unnecessarily unwieldy.[6] The use of an even-number scale forced reviewers to make clear decisions by avoiding the choice of a middle point.

2.5.5 Standardisation of CRAS judgements

In studies 1a, 1b and 2a, there was no attempt beyond the briefing and discussion of the four factors to standardise reviewers' judgements. However, for study 2b, the lead reviewer had prepared an additional briefing, using a number of selected questions to try to bring the team to a shared understanding about the application of the numerical scale. This pilot attempt to standardise reviewers' criteria for making judgements was well received by reviewers, who found that it increased their confidence in making the numerical judgements. It certainly seemed to have had a positive effect, in that there were

---

[5] These factors had been identified in a study into question structure by University of Cambridge Local Examinations Syndicate (UCLES) commissioned by QCA. Each factor has the capacity to make examination questions more or less difficult, irrespective of the subject content. The exact interpretation of the four factors is often, to a degree, subject dependent. Explaining any subject-specific aspects was one of the tasks carried out by the lead reviewers.

[6] In effect, the scales used in all the studies allowed a four-point scale for each level with a two-point overlap across levels, 1–4, 3–6 etc. At the same time, it allowed room for judgements to reflect cases where reviewers felt that particular papers were not set at the appropriate level.

no significant differences of opinion between reviewers about the particular numerical ratings.

2.6 Instruments and materials for the review of candidate work

The review of candidate work had two distinct functions. The main aim was to consider the standards of performance required to achieve a given level of attainment in a subject. It also provided a key opportunity for reviewers to consider the judgements they had already made about the demands of the examination materials. This was important because it was not always easy to judge precisely the demands these materials made. One of the key causes of this was that many mark schemes made use of relative terms such as ‘sound’ or ’effective’. The examiners' interpretation of these terms was impossible to determine without seeing examples. Thus reviewers needed to see marked scripts in order to be confident that their own interpretation was the same as that used by the examiners.

There was a second reason why allowing reviewers to see candidates’ work was important. The reviewers were, by definition, subject experts. However, those taking the papers were, to a large degree, novices. It is a commonplace of examination experience that candidates find questions, and sometimes whole papers, much harder or easier than those setting them have expected. The examples of marked work allowed reviewers to evaluate their perceptions of the tasks in terms of how the candidates actually responded to them.[7] For study 1a, this proved to be particularly revealing.

For the studies, awarding bodies supplied candidates’ work based on the same specifications as those used for the review of examination demand. It was decided not to use coursework in these studies. There were several reasons for this. First were practical considerations. It is very difficult to obtain candidates’ coursework, since it is held in the schools and colleges and, indeed, often returned to the candidates themselves. Coursework is also often very bulky and extensive, which greatly affects the range of candidates' work that can be scrutinised in a given period of time. In addition, previous experience using coursework in such exercises tends to suggest that it brings in so many other variables that that they would interfere with the particular aim of these studies, which is quite complex enough. The candidates’ work for this exercise therefore

[7] So important is the expert/novice aspect that for studies 2a and 2b QCA piloted an exercise whereby able students taking the relevant subjects were asked to comment on the materials. The exercise was on far too small a scale to be reported here, but was very encouraging. It is hoped that the methodology can be developed so that it can be used in future research into comparability.

comprised the whole externally assessed work of candidates. The absence of coursework, which is an area where the assessment is most dependent on potentially differing interpretations of assessment criteria, does however affect the confidence that can be placed in the findings. This is particularly relevant where the proportion of coursework differs significantly between the subjects under review, as in study 2b.

The differing nature of the studies meant that there were differences in the nature of the materials used. For study 1a, which was trying to look across levels of the National Qualifications Framework, the work came from candidates whose performance fell roughly in the middle of each level. This had the advantage of meaning that it was not of a standard that reviewers were used to from grade awarding committees, which focus on the thresholds for grades A, C and F at GCSE and grades A and E at AS and A2. It was hoped that, as a result, they would not be distracted by trying to decide whether the grading standards were correct, focusing instead on whether they were comparable. This is consistent with the methodology used for making the comparisons described below. Feedback from the expert group resulted in a slight change for study 1b. Here the work was still from the middle of the grade range covered by the level of the qualification system, but it covered a range of attainment. (In study 1a, the work came in principle from a specific point in the range.)

For studies 2a and 2b, it was felt that one of the key questions that the investigations should to try to answer was whether the grading standards were consistent across the subjects. Such standards reside largely in the key judgemental grades determined by awarding committees: grades A and E at AS and A2. For these studies, therefore, the candidates were those who had just gained an E or an A in their examinations.[8]

2.7 Process for review of candidate work

The review of scripts took place at residential meetings convened for the purpose. Reviewers worked independently, making an extensive series of comparisons over two days. Reviewers worked in sessions of approximately one hour. For each session they were provided with two piles of candidates' work, one from each subject, at the same grade and qualification level. They worked through the piles drawing comparisons

[8] In fact, on analysis it became clear that not all the work provided conformed closely to the specification for the study. As a result, the work covered a known range of attainment close to but not exactly corresponding to borderline performance. Although this could be taken into account to an extent during analysis of the outcomes, it inevitable reduces the confidence that can be placed in the findings.

between the performance of individual candidates, typically making at least nine comparisons per session. For each comparison they were asked to decide which candidate's performance was better; no 'ties' were allowed. They recorded their decisions on forms that also provided a space for any comments about which aspects of the candidates' performance had contributed to their decision. This was to help to identify whether there were aspects of performance that reviewers tended to value highly or place less value on. It also identified differences in the ways subjects tended to draw out performance and reward it.

When the results of the many comparisons made are aggregated and analysed, it is possible to see whether or not performance in one subject is judged by reviewers to be better than that in another, given that the candidates have received the same grade.

2.8 Evaluation

In all phases of the studies participants were encouraged to provide comments upon the process and make suggestions for improvement. Many availed themselves of this opportunity and their feedback proved very helpful in refining and amending the process. In addition, QCA has convened a group of experts in assessment research to comment and advise on its work in monitoring of standards and comparability. This group also provided a number of important suggestions for amendments to the methodology and on the interpretation of the outcomes.

3. Findings of the reviews of examination materials

3.1 Study 1a: GCSE, AS and A level geography and history

Reviewers assessed the specifications in terms of several factors. Their key conclusions are presented below.

Schemes of assessment

At GCSE, the schemes of assessment were similar in many respects, though the emphasis in geography was on a problem-solving, issues approach, compared with the linear approach and literate emphasis of history.

At A level, the two specifications had different approaches to the design of the AS component. Geography aimed to ensure coverage of broad areas of knowledge, understanding or skills. In contrast, history offered and appeared to encourage narrowness of study.

The geography specification had a clear focus on the inter-relationship between people and their varied environments, and the issues related to management arising from those relationships. The history specification had no specific historical rationale; rather, it placed great emphasis on providing the opportunity for teachers to construct their own course from the available options.

Subject content and options

Geography had greater breadth and a balance of content than found in history across GCSE, AS (in terms of compulsory content) and A level. History offered a range of options and therefore demand depended on choice of options by centres, which, in turn, carried the potential for narrow historical content. There was also variation in demand between optional routes. In geography, questions were often structured to provide an incline of demand. This contrasted with the history assessments, where differentiation was usually by outcome, and in which questions were generally more predictable and familiar in style and content.

At A level it was difficult to judge the depth of knowledge and understanding required for both specifications, though the mark scheme for geography appeared to be less

demanding. The top mark band used ‘sound’ as a qualifier rather than ‘comprehensive’, used in history.

Nature of assessment materials

Reviewers judged that the language demands of the history assessments at all levels were much greater than those for geography. Geography questions were more structured, with more accessible language, and there was use of tiering at GCSE. History papers were more open-ended, with essay-style questions requiring considerable intellectual and communication skills to structure a logical response, with a greater emphasis on literary demands and quality of written communication. At GCSE and AS level, however, the history questions also tended to be repetitive in style and to require large amounts of recall.

There was a wider variety of assessment tasks in geography than in history. Geography required a greater ability to respond to a variety of tasks and to demonstrate a range of enquiry, organisational and communication skills. Reviewers recognised that these tasks at A level required complex preparation, independent working and use of analytical strategies.

At GCSE, the resource-based questions in history required interpretation and analytical thinking and were very challenging for the whole ability range. In geography the resource-based questions were demanding for foundation tier candidates.

As already noted, the GCSE history papers rely largely on differentiation by outcome, whereas in geography differentiation is by task. The fact that geography is tiered allows for even greater targeting of the tasks.

Outcomes of CRAS analysis

Reviewers also carried out a CRAS analysis of the question papers. The results are summarised in Table 3.

Table 3: Average ratings arising from the CRAS analysis of geography and history question papers, by level

	Geography	History[9]
GCSE foundation tier	3.1	5.6
GCSE higher tier	4.3	
AS units	6.3	7.3
A2 units	8.5	9.1

Key to expected ranges	
GCSE foundation	1–4
GCSE higher	3–6
GCE AS	5–8
GCE A2	7–10

From Table 3 it can be seen that in general the papers did follow the expected progression up the scale. It can also be seen that the history papers were consistently judged to be rather more demanding than the geography ones. It was particularly striking that the average ratings for the untiered GCSE history papers were higher than for either of the GCSE geography tiers, even the higher tier papers. In fact, the mean rating for the GCSE history papers was not far below the expected maximum for GCSE, suggesting that reviewers considered the papers very demanding. It is important to note, however, that these views were considerably revised during the script review.

Coursework

At GCSE, reviewers judged the complexity and demands of coursework to be similar in both subjects.

[9] GCSE history question papers are not tiered.

Progression

Reviewers judged that the content of both specifications at AS was 'about right' and sat comfortably between GCSE and A level in demand. However, they considered that the open-ended essay questions in history were more demanding than the short structured questions with low mark tariffs in geography, and that overall the incline of demand from GCSE to AS was markedly steeper for history than for geography.

Synoptic assessment at A level

The approach to synoptic assessment was different in the two subjects. In history, questions were set on specific periods of history, which may differ from previous periods studied. They did not require the demonstration of knowledge and understanding of connections across other parts of the specification but they did require high levels of skill to be demonstrated in this new knowledge context. The geography synoptic paper had less emphasis on new knowledge but did require high-level thinking and analytical skills, and the ability to draw on understanding from other parts of the specification.

Overall findings

Overall, variations in the schemes of assessment were considered a more important influence on variations in demand than the intrinsic characteristics of the two subjects.

The GCSE history assessment was judged to be more demanding than geography, especially for foundation candidates.

At both GCSE and AS, history was judged to be more demanding than geography, but both subjects were considered to be in line for the A level overall.

Reviewers considered that at A level the specifications had very different assessment characteristics, with an emphasis on different assessment objectives, but that they were of similar demand. However, with their different characteristics, the demands would be different for candidates of different aptitudes.

In every case except geography at AS, the ratings suggested that the subjects were slightly too demanding, with GCSE history being seen as significantly demanding.

3.2 Study 1b: GCSE, AS and A level sciences

Assessment objectives

Common assessment objectives were used across the sciences at GCSE and A level. Although reviewers recognised the rationale for using common assessment objectives, they found that this approach did not reflect the individuality of each subject. In addition, they judged that it did not actually ensure comparability, as the impact of the assessment objectives varied between the subjects.

Schemes of assessment

Reviewers judged that the schemes of assessment for the separate sciences were significantly more demanding than that for double award science at both tiers of GCSE. They found physics at higher tier the least demanding of the three separate sciences in this respect.

Reviewers found the demand of AS to be generally about right, despite differences in examination time across the subjects. At A2 chemistry was judged more demanding than biology or physics, although the approach to synoptic assessment in chemistry was not considered as demanding as that of the other two sciences.

There were differences in the way choice was used across the subjects at A level. The physics specification offered some options at AS, whereas biology offered choice only at A2, and this had an impact on demand. Whether the options were found in AS or in A2 was judged to have a differential effect on demand.

There were differences between the subjects in the assessment of the ability to synthesise knowledge, understanding and skills (assessment objective 4). This had a major effect on demand. In chemistry[10] a multiple choice / completion test was used to test knowledge recall across the course, but this approach required little analysis or synthesis and reviewers judged that it was less demanding of synoptic skills than the approaches adopted in the other two sciences. Physics and, to a greater extent, biology had more demanding ways of assessing assessment objective 4, often through extensive writing. In addition, the allocation of 40 per cent of A2 marks to a two-hour synoptic paper in physics significantly increased the demand.

[10] This was a particular feature of the specification used in the study. In other chemistry specifications, different approaches are used for assessing synoptic skills.

Syllabus content

Reviewers identified several distinct problems with syllabus demand, noting in particular that there was a heavy content load in all subjects both at GCSE and A level. Chemistry was judged to be very demanding in terms of content at all levels.

Reviewers identified the following issues at GCSE in all four subjects (to a greater or lesser extent).

- A particular piece of content within a topic was sometimes assigned to the higher tier when it could have easily rested in foundation tier and made for more coherent teaching and learning. As a result, questions relating to the topic at foundation tier were narrow in their focus.
- Where a whole topic area was assigned to foundation tier, large elements of the topic were often demanding for the candidates. Conversely, where a whole topic area was assigned to higher tier, large elements of the topic were often undemanding for the candidates.
- There were serious reservations about the usefulness of foundation tier subject content. Reviewers did not consider it relevant for those not going on to further study.
- Some areas of subject content were conceptually very demanding. They were perhaps too abstract, lacking in relevance to everyday life or low in pedagogical opportunity.
- Overall, reviewers considered the level of demand of the syllabus content in chemistry to be too high for both foundation and higher tier.

There were also some concerns specific to AS and A2.

- The level of demand in all three subjects was unnecessarily high.
- Chemistry was more demanding in content terms when compared with physics and biology.
- Optional routes led to differences in demand within a subject.

Reviewers noted the inter-relationship of the sciences and their relationship to mathematics. This was particularly the case for progression from AS to A2. It would be difficult, for example, for a candidate to succeed in biology at A level without a confidence and competence in chemistry, as well as some mathematical knowledge. Similarly, A

level physics had less overlap with the other sciences but required fluency in mathematical and symbolic thinking, in addition to a knowledge of specific mathematical techniques.

Nature of assessment materials

At each level there was variation in the types of questions used across the subjects. For example, biology papers tended to require more extended writing, whereas in physics the most common task types were calculations and short explanations, with calculations used especially in more demanding questions.

The range of question types within subjects was very limited, particularly in chemistry, where there was little variation from short or structured questions that required a high amount of recall.

Reviewers judged that physics tested applications more thoroughly than the other sciences. They found that there was a demanding content load in biology. Reviewers considered that the assessment of evaluation was limited in all sciences, as it took the form of interpreting data from tables and graphs.

Overall at GCSE, reviewers judged that papers did not differentiate successfully, with too little that would be accessible to the weakest foundation tier candidates or challenging to the best higher tier ones. They considered that this was exacerbated by the lack of direct assessment of key stage 3 material.

The assessment of the quality of written communication was inconsistent across subjects. For example, there was little extended writing in physics, which meant that the quality of written communication could not easily be assessed. In addition, there was little detail with regard to the assessment of the quality of written communication, so that the standards applied were unclear.

Subjects adopted different approaches to open-ended questions, leading to probable differences in demand. These differences lay in the amount of extended writing required, in the nature of the tasks set and in the expectations laid out in the mark scheme.

Reviewers found that the time allocation for some of the papers was insufficient. In particular, the chemistry papers and the AS physics topics test were very demanding in terms of the time available.

In addition, reviewers noted that demand relating to mathematical work was variable. While biology required relatively little quantitative thinking, in physics complex

mathematical processes were repeated unnecessarily and used in place of writing and synthesis to increase demand.

Outcomes of CRAS analysis

Reviewers also carried out a CRAS analysis of the question papers. The results are summarised in Table 4.

Table 4: Average ratings arising from the CRAS analysis of biology, chemistry, physics and double award science question papers, by level

Subject	Biology	Chemistry	Physics	Double award
Foundation	2.5	2.6	2.9	2.5
Higher	4.1	4.5	4.5	4.3
AS	6.7	6.7	6.8	N/A
A2	8.2	8.4	8.2	N/A

The table shows that all subjects were similar in demand at each level, with chemistry and physics perhaps slightly more demanding than biology and, at GCSE, double award science. That they were very similar in terms of demand was unsurprising since all three subjects adopt a very similar approach to assessment, with the majority of questions requiring short answers. A series of structured questions is often employed.

This table shows the ratings averaged across the four CRAS factors. It therefore doesn't reveal any variations between them. In this context, the analysis of chemistry at all levels, generally yielded high ratings for ‘abstractness’, often with ‘complexity’ and ‘strategy’ lower. Where ‘strategy’ was high, the questions required considerable organisation and, for example, the selection of information from the data sheet. Calculation questions generally did not provide strategies for undertaking the calculations. In addition, the ‘resource’ requirements for chemistry at AS and A2 were high when compared with those of the other sciences. This is reflected in the comments elsewhere about the high amount of recall required in chemistry.

Demands in relation to coursework

At GCSE the demands of coursework were broadly equivalent across all the subjects, though some reviewers found that double award science was more demanding than the separate sciences because there were two marks per skill area. It was also suggested that the coursework requirements were poorly matched to biological investigations (with so many uncontrollable variables, the difficulty of finding quantitative opportunities, the time over which data must be collected and the need for controls).

At AS and A2, the coursework criteria were natural developments of those at GCSE. Reviewers judged their demand to be about right in biology and chemistry. In physics[11], however, the practical test offered a very narrow interpretation of the assessment objective. In addition, very little was required from candidates in terms of devising strategies in the examination; rather, instructions were given which candidates had to follow step by step. The complexity of the examination was also much less than the complexity of completing coursework tasks. Overall, the level of demand for physics in this area was lower than it should be.

Progression

Overall, reviewers found that progression from GCSE to AS and from AS to A2 was not equivalent across the sciences and that this may have been linked to the variable impact of the assessment objectives across the subjects. Reviewers judged that the step from GCSE double award to all three separate sciences at A level was particularly large.

Summary

- At each level, chemistry was seen as the most demanding in terms of content.
- In all subjects at GCSE there were oddities in the way content had been divided up between the tiers.
- Reviewers expressed reservations about the usefulness of some foundation tier subject content. Reviewers did not consider some content to be relevant to those not going on to further study.

[11] This was a particular feature of the specification used in the study. In other specifications, practical skills can be assessed through coursework, rather than through a compulsory practical examination.

- Reviewers considered that successful progression from AS to A2 depended on knowledge of other sciences and, to a varying degree, of mathematics.
- Options within subjects affected the demand. Some options were conceptually harder and/or tested in more a demanding fashion. The differences between subjects in terms of whether they offered options at AS or A2 affected both comparability between subjects and progression within a subject.
- At A level the approach to the assessment of the ability to synthesise knowledge, understanding and skills (assessment objective 4) varied between the three sciences, with chemistry judged to be less demanding than physics and biology in this respect.
- There was variation in the range of task types used across the subjects at both GCSE and at A level, with biology tending to require more extended writing and physics using more questions requiring short answers or calculations. However, the range of task types within the subjects tended to be limited and this was particularly the case for chemistry, which tended to use questions requiring short answers, with a high amount of recall.
- Progression from GCSE higher tier to AS, and from AS to A2 was not comparable across the subjects.

3.3 Study 2a: A level biology, psychology and sociology

Assessment objectives

Reviewers considered that in spite of the differences in the number of assessment objectives in different subjects, overall they were of similar demand.

Subject content

Reviewers judged that, while there were clear demands for previous subject-specific knowledge in biology, but not in psychology or sociology, this did not necessarily impact on the overall demand of the subjects.

Reviewers did express concern that, in the case of sociology, it was possible to answer almost all of the two-mark questions on the basis of non-specialist knowledge. This was not the case for either biology, where even short-answer questions were judged to be very demanding in terms of the specific subject knowledge required, or psychology, where students could use knowledge of everyday situations in their responses but had to make clear the link to psychological principles in order to receive credit.

In terms of the content to be covered, the biology specification was judged to be the most demanding, given the combination of intellectual and practical skills required. Further, the nature of the question papers, with all questions being compulsory, meant that candidates had to cover all content in appropriate depth and detail. Psychology and sociology were judged to be less demanding in terms of the volume of content to be covered, and it was thought that the fact that candidates had some choice in the questions they answered might reduce some of the pressure to cover all the content in depth and detail. In comparison with sociology, reviewers judged that psychology candidates would need to demonstrate a wider and deeper knowledge of subject-specific principles to gain credit at AS level. In particular, the level of choice offered in sociology had the potential to lead to a rather narrow course of study.

Overall, reviewers judged that the content of the sociology and psychology specifications were appropriate for AS level. They judged that the content of the biology specification was very demanding and that this demand was increased by the fact that, across AS papers, there was no question choice.

The nature of the assessment materials

Reviewers considered the style of assessment. They judged that, in terms of the language used, the biology and psychology specifications could be very challenging for candidates. Many technical terms were used and candidates were expected to be familiar with them. In the case of sociology, at AS level there was less emphasis on technical language.

The approaches to assessment in both biology and psychology guided candidates in their selection of material for a response. In the case of biology, it was judged that, as long as candidates had learned the necessary material, the questions should enable them to demonstrate their knowledge and understanding. In the case of sociology, there was concern about some of the short-answer questions demanding no more than comprehension of a passage. Overall, however, it was judged that the questions should enable candidates across the ability range to demonstrate their knowledge.

At A2 reviewers expressed concern about the possibility of candidates offering and being given credit for GCSE knowledge in responses to A2 biology questions, whereas candidates responding to psychology and sociology questions on A2 papers would be addressing unfamiliar material. This potential problem was ameliorated, however, by the very high demand of the remainder of the A2 biology content. Similarly, concern was raised about the possibility of sociology candidates being able to repeat material from the AS unit 3 in the A2 unit 5, which would reduce the overall demand of the assessment. Overall, however, reviewers judged the specification content demands for both sociology and psychology to be appropriate for A2, but found those for biology to be high.

Reviewers judged that the biology question papers were very demanding. While the short-answer structure of many of the biology questions was thought to be helpful to candidates, the high level of very detailed and specific subject knowledge required, combined with a prescriptive mark scheme and the fact that all questions were compulsory, made the biology papers very demanding indeed.

In terms of accessibility, sociology and psychology question papers were judged to be broadly comparable, with clear attempts having been made to ensure a range of question types to allow candidates from across the ability range to access the papers and demonstrate their knowledge and skills. The more open nature of the questions and the mark schemes in both sociology and psychology meant that candidates were able to select their own material to use when responding. Therefore, sociology and psychology could be judged to be less demanding than biology, as candidates could, in some sense,

hide their ignorance. It was judged, however, that weaker candidates would be unlikely to be able to do this and, in the case of psychology, the stringent demands of the mark schemes for candidates to couch their responses in appropriate psychological terminology and theory, should prevent this. On the basis of the proportion of marks available for non subject-specific knowledge and comprehension in the case of data-response questions, reviewers judged that the sociology question papers and mark schemes were the least demanding.

In terms of the language demand of questions and, where appropriate, source materials on question papers, biology and psychology were judged to be broadly comparable in terms of demand. Biology students were found to have to deal with more demanding numerical and graphical material, while psychology students had to deal with a range of materials of different types, with a high technical language demand. Sociology question papers were judged to be marginally less demanding at AS, although broadly comparable at A2.

Sociology and psychology both made heavy demands on candidates in terms of language, with a mixture of short answers and extended writing required. This made both subjects demanding in terms of candidates' ability to select information and organise ideas. In biology, there was no great emphasis on the skills of extended writing, even in the case of essay questions, where the focus of the mark scheme was on specific points of content, with relatively little emphasis on the quality of written communication.

Outcomes of CRAS analysis

Reviewers also carried out a CRAS analysis of the question papers. The results are summarised in Table 5.

Table 5: Average ratings arising from the CRAS analysis of biology, psychology and sociology question papers, by level

	Biology	Psychology	Sociology
AS units	2.6	2.8	3.1
A2 units	2.9	4.4	4.2

From Table 5, it can be seen that there was very little difference between psychology and sociology at either AS or A2. It can also be seen that both were judged as significantly

more demanding than biology at A2 and a little more demanding at AS. In particular, there was very little difference in the ratings for biology between AS and A2.

It is certainly true that there was little difference in the approach to assessment taken for biology at AS and A2. The difference lay in the range and complexity of the subject content. Reviewers judged that it was difficult to make comparisons between the biology question papers on the one hand and the sociology and psychology papers on the other. The judgements they made certainly reflected a real difference in approaches to assessment across the subjects in terms of the cognitive demands they made. However, they considered the distinction misleading in that it would not truly reflect the difficulty faced by candidates in completing the papers. The same distinction lay at the heart of the review of history and geography in study 1a.

Coursework

Reviewers noted that there were potentially significant differences between the subjects in terms of coursework. Reviewers judged that the combination of independent research skills in coursework and the assessment of practical skills made biology slightly more demanding than sociology and psychology. There was also some concern that the high level of guidance given to psychology candidates in the specification might reduce the demand of the coursework option. However, the need for candidates to demonstrate analysis and evaluation in both qualitative and quantitative contexts was demanding. The coursework demands for sociology were judged to be appropriate.

Optionality

For biology candidates, with the exception of some choice in the essay question, all questions on all papers were compulsory. Sociology and psychology candidates could choose a route through the papers. Concern was raised about the possibility of both sociology and psychology candidates having studied a relatively narrow range of content, depending on the route through the specification selected by the particular centre.

Time

Reviewers judged that, although candidates for biology had to answer a considerable number of questions, the relatively small requirement for extended writing meant that the time available for each paper was appropriate. The same was the case for sociology, which, although it had a greater reading demand in terms of volume than either biology or psychology, gave candidates sufficient time. Psychology was judged to be the most demanding in terms of time pressure per question.

Synoptic assessment

In the case of both psychology and sociology, the synoptic units required candidates to make links between different aspects of the course. However, reviewers were concerned that the level of choice in the non-synoptic units meant it was possible to select a relatively small range of subject areas and, effectively, ignore others. This meant that candidates could receive credit for demonstrating knowledge of the subject as a whole, while not addressing important aspects. Further, because the synoptic unit had to be accessible to candidates who had studied a range of different units, the questions tended to be generic. Reviewers were concerned that these generic questions would become very predictable and lead to prepared responses.

In the case of biology, however, the synoptic unit was judged to make appropriate demands in terms of the coverage of material from the non-synoptic units. This material had to be selected and applied in a given context, and this was judged to be demanding for candidates.

Progression from AS to A level

The CRAS analysis showed that the biology papers were not significantly more demanding at A2 than at AS. In terms of the structuring of the questions, there was very little difference between AS and A2, but reviewers judged that this was balanced by the demand of the content. Thus the outcomes of the CRAS analysis need to be seen in tandem with the nature of the subject content. Reviewers argued strongly that the issue was not that biology was less demanding than sociology or psychology at A2, but that it was probably rather too demanding at AS, where its relatively similar demand in terms of CRAS rating made no allowance for the very demanding content.

In the case of sociology and psychology, the demands of the A2 assessment materials were significantly higher than the demands of the AS assessment materials, both in terms of the structuring of the questions and the demand of the content. At both levels, the outcomes were broadly in line.

Overall specification and assessment materials comparison

Overall, reviewers judged that biology demanded the greatest breadth of detailed knowledge, requiring students to demonstrate high levels of recall. It did not, however, make the same demands in terms of evaluation/interpretation as sociology or psychology.

Sociology was judged to be potentially very demanding, because of the requirement to contextualise judgements in appropriate theory. There was concern, however, that non-contextualised, common sense responses could receive too much credit, and that the most popular route through the AS would give candidates too much opportunity to write uncritically, from their own experience. The lack of prescription and apparent leniency of the mark schemes added to overall concern.

Psychology did not give candidates credit for anecdotal knowledge, but, instead, the mark schemes made clear demands for candidates to use correct technical terminology and to couch their answers clearly in appropriate psychological theory. Psychology was judged to be technically demanding.

In the case of psychology and biology, the question papers made use of complex concepts. Candidates could neither infer nor guess answers and this rigour was maintained by demanding mark schemes. This was not clearly the case in sociology.

3.4 Study 2b: A level English literature, history and media studies

AS comparison

Overall, the history assessment materials were judged to be slightly more demanding than those in media studies. In all the units in English literature there was variability in demand both in set texts and the questions asked about them, which made it very difficult to make a judgement about the relative demand of the subject compared with history and media studies.

The history units required significant levels of selection, analysis and synthesis, all drawing upon knowledge of the chosen historical periods, to support argument and evaluation. The media studies units required candidates to demonstrate skills of textual analysis and some ability to contextualise that analysis within broader cultural issues.

In terms of depth, the demands of the history assessment materials were judged to be greater than those of the media studies assessment materials. This was because of more stringent demand for detailed knowledge and sustained conceptual analysis in the history mark schemes. However, in terms of breadth, the media studies assessment materials were judged to be more demanding, as they required the development of the skills of practical production as well as those of conceptual analysis.

The English literature specification moved from textual analysis towards contextualisation, but this was judged to be somewhat uneven, and so the overall demand was affected by variability according to the choice of text and examination question. There were some texts that were judged to be particularly inaccessible, and this could result in the candidate first having to engage with a very difficult text and then having to engage with a very demanding question. In all the English literature units the demand of some questions was judged to be too high for AS.

There was some difference of view among reviewers as to whether the relative accessibility and familiarity of the type of textual material found in the media studies assessment materials was an advantage or a disadvantage to candidates. Some reviewers argued that the relative familiarity of the media studies material could put candidates at an advantage. Further, some of the materials were potentially very enjoyable and so motivating for candidates. Others argued that the familiarity could predispose candidates to respond in an inappropriate and possibly colloquial way to questions, and so fail to demonstrate their skills.

A level comparison

Judgements about comparative demand at this level were made difficult by the very different structure of the assessment objectives across the three subjects.

In history there were common assessment objectives at both AS and A level and so, while there was some evidence of relatively high demand at AS level, when the A level was taken as a whole, it was not judged to be more demanding than the other two specifications.

In English literature, there were some common assessment objectives across the levels, but also some that were specific to AS level and some specific to A level. The level of demand in English literature was judged to be greater at A level than at AS level and this was judged to be entirely appropriate. There was concern, however, that there was variation in demand at both levels according to question choice, and that this derived from the different ways in which the assessment objectives requiring knowledge of the contexts of the literary texts (assessment objectives 5i at AS and 5ii at A level) were interpreted in particular examination questions.

The assessment objectives for the media studies specification were so arranged that each unit was designated a specific and discrete assessment objective.[12] This meant that the respective demands of the two levels were clearly differentiated, although the principles of progression and coherence between units were not so clearly mapped as they were for history and English literature. For this specification, the most significant increase in demand, in terms of the presentation of conceptualised and evidenced arguments on unprepared questions, was found in the final unit.

Nature of the assessment materials

There were some types of task that were similar across all three specifications. Continuous prose answers and essays featured heavily in all three subjects. All the English literature tasks required essay responses, although there were some instances in which questions were supported by prompts about the key areas to be included within the answer.

All three sets of assessment materials required candidates to respond to unprepared material presented on or with the examination paper. In media studies one of these

[12] This is not true of other awarding body specifications in A level media studies. There must therefore be questions asked about the extent to which these findings can be generalised.

sources was a video extract. In both English literature and history there were tasks that required comparison, synthesis and evaluation across a range of unprepared source materials.

The assessment activities were judged to be broadly comparable across all three specifications. While some individual tasks within media studies represented rather lower demands for synthesis of different sources, this specification had, overall, a greater variety of types of conceptual and technical demand, including a significant weighting (40 per cent) for practical production. This breadth was judged to compensate for the relative lower demand of the assessment activities in media studies. The range of skills that candidates for English literature were required to demonstrate was judged to be narrower than for media studies or history, but this was judged to be balanced by the relative abstractness of many of the tasks and the degree of selection required from increasingly demanding texts, in presenting extended arguments.

Some concern was expressed about the extent to which the specification and assessment materials for media studies allowed candidates to use prepared material. This concern was borne out at the script review, where there was some evidence of candidates drawing on prepared responses.

Outcomes of CRAS analysis

Reviewers also carried out a CRAS analysis of the question papers. The results are summarised in Table 6.

Table 6: Average ratings arising from the CRAS analysis of English literature, history and media studies question papers, by level

	English literature	History	Media studies
AS Units	4.3	4.1	3.4
A2 Units	4.7	4.6	4.5

As the separate figures for AS and A2 imply, media studies assessments provide a much steeper incline of difficulty between AS and A2 than either English literature or history, with the three subjects much more closely aligned at A2.

Content coverage

The content coverage of all three specifications was judged to be suitable and broadly comparable.

The opportunity for breadth of study was judged to be greatest within the history specification, but, in view of the wide range of optional areas of study within this specification, it was possible for students to focus more narrowly on periods and topics, building knowledge incrementally rather than studying a particularly wide range of content. For this reason, judgements about demand, in terms of content coverage, were difficult to make for history, as they were so dependent on choice of route.

The content coverage for English literature was, for similar reasons, potentially very demanding. However, the demand was influenced by the choice of texts and it was possible for candidates to select a route through the specification that focused on less demanding texts. Two of the Shakespeare texts could have been studied at earlier levels, and although the specification points out that candidates should select a text that they have not studied previously, it was not clear how this requirement could be enforced. Therefore, it remained possible for candidates to select a more familiar and so potentially less demanding route through the specification. Further, while the study of literary, historical and social contexts in English could be very wide indeed, the number of optional questions allowed centres to focus on relatively narrow interpretations of the context and for candidates to select questions in line with these narrow interpretations.

The amount of choice within both the English literature and history specifications and assessment materials, compared with the relative lack of choice in the media studies specification and assessment materials, gave rise to concern about making judgements about the comparative level of demand. Overall, however, it was judged that the apparently heavy demands, in terms of breadth and depth of the English literature and history specifications and assessment materials, were balanced by the high level of choice of content and examination question, and so were broadly equivalent to the relatively lower demands, in terms of depth, but higher demands in terms of breadth and lack of choice in the assessment materials in media studies.

Optional routes

As stated above, the history specification allowed for a very large number of optional routes, in terms both of topic areas and alternative questions. English literature had a coursework/written examination option in unit 5 only, although there was considerable choice of texts for study within the individual papers. In media studies, there was only one

route through the specification, though there were optional areas of study within the overall structure and considerable scope within the two coursework productions for candidates to develop particular interests and areas of study. It was judged that, given the relative demands of the content, this variability in the degree of optionality did not impact on overall demand.

Where options were offered in the examination papers for history and media studies, there appeared to be only minor inconsistencies in demand. This was less clearly the case for English literature, where uneven demand was a concern.

Synoptic assessment

The quality of synoptic assessment was judged to be very sound across all three specifications, with the demands of the English literature synoptic unit judged to be very high. Reviewers commended the approaches taken in all the synoptic units and the extent to which these units represented broad, thematic approaches to the subject, integrating knowledge and skills developed throughout the course.

Overall specification and assessment materials comparison

It was judged that, overall, there was no clear evidence of significant differences in demand between the three specifications.

4. Findings of the review of candidates' work

4.1 Background

Before presenting the outcomes of the review of candidates' work, it is important to remember certain important points. The first is that this was the aspect of the work that most changed over the course of the various studies, both in terms of the selection of work and in terms of how it was analysed. It would be particularly dangerous, therefore, to try to look across any of the studies and draw wider conclusions.

The second important point is that there were key deficiencies in the work considered. In the first place, only performance in external written examinations was reviewed. This is likely to have been a factor in each study but is, naturally, likely to have been more important in subjects where the coursework elements were judged to be significantly different, whether in volume or demand. It is also important to note that in this context different weighting for coursework sometimes meant that reviewers were trying to compare performance across different numbers of components. For example, in study 2b, the work reviewed did not include candidates' coursework in media studies. This work was worth 40 per cent of the overall A level marks and involved significantly different skills to those required in English and history.

The judgements that reviewers were being asked to make were very complex in any case, much more so than in any comparability study looking only at standards within a subject. There is plenty of evidence that it is very hard for experts to make reliable judgements about work, even when looking at the performance of different candidates in the same subject. The whole awarding process is structured to take this into account, with judgements on candidates' work used to identify a zone of uncertainty, which may be several marks wide, within which a grade boundary will lie. A range of other information is then brought in to enable a recommended grade boundary to be reached.

Given this, although the kind of analysis that is carried out in reviewers' judgements is capable of producing a high level of accuracy, it seems likely that such accuracy would be specious. In this report, and the separate reports on each study to be published later, only cases where differences in standard were large are reported. Even here, as the outcomes of study 1a highlight, reviewers were clear that the substantial differences identified did not necessarily arise from differences in standard between the two subjects; instead, the results caused them to reconsider their judgements significantly about the actual difficulty of the examinations for the candidates. In many ways, this may be seen as the most significant finding of the work, in that it raises real questions about the most effective way

to assess candidates' knowledge and understanding, and in particular, about effective differentiation.

4.2 Outcomes of study 1a: GCSE, AS and A level geography and history

The work reviewed was drawn from middle point of the range of attainment covered by each level, that is, the E/F boundary for foundation tier GCSE, the A/B boundary for GCSE higher tier and mid C for both AS and A2. The work comprised the complete examination work of candidates.

At GCSE, for candidates with grades in the middle of the foundation tier (about grade E), reviewers were overwhelmingly of the view that the history candidates performed better than those in geography. For those in the middle of the higher tier (about grade A), the view was slightly less consistent but still strongly favoured the history candidates. At about grade C at AS, the balance of judgements were still more even, but clearly suggested that the history candidates were stronger. Only at A2, again using work from the middle of the grade range, was the position reversed, with the geography candidates judged to be slightly stronger.

However, as already suggested, the reviewers were strongly of the view that these findings did not mean what they seemed to suggest. Essentially, they traced the issue rather to the nature of the assessment. It is notable that it was only at A2 that geography candidates were required to write significantly extended answers of the type that was expected of history candidates in all examinations. In fact, the more similar the assessments became, the closer the balance of judgements was. GCSE geography papers at foundation tier consisted almost entirely of questions made up of short parts calling for very specific pieces of knowledge and understanding from the candidates; at higher tier it was much the same, but there were a few opportunities to provide more extended explanations. At AS, questions were still often highly structured and very specific in the information required. Only at A2 was there a large element of open-ended questions requiring essays and other forms of extended answer. Notably, at this level the geography candidates were judged to be slightly better.

In history, the papers were more structured at AS than at A2 and more structured still at GCSE. However, they remained much more open in terms of the answer expected and, even at GCSE, called for some quite extended writing. What reviewers found was that the relatively open nature of the history questions was enabling for all candidates and most significantly so for the foundation tier candidates. In other words, the untiered history examination was a more effective differentiator than the geography examination, which had papers specifically targeted at lower and higher attaining candidates. Moreover, the

history papers allowed candidates across the grade range to demonstrate their knowledge and understanding. This shows that they were also effective discriminators.

Reviewers commented that the range of task types used in history papers at each level was relatively narrow, with most questions being open-ended and requiring an extended written response. This meant that history papers were very accessible for candidates with good literary skills who performed well at this type of task. The papers also allowed candidates to conceal significant gaps in their knowledge. This was clearly not the case in geography, where candidates had to cope with a greater range of task types and where, in many cases, they had to demonstrate very specific knowledge.

These findings raise some important questions about the methodology. The first is whether it is appropriate to look across subjects, especially if the approach to assessment differs significantly. In fact, the reviewers were of the view that the exercise, though difficult, was possible and that their judgements were valid. The key thing, they argued, was how those judgements are interpreted. At a deeper level, it does bring home an important issue about all comparability studies: that those making the judgements are essentially experts; those who take the examinations are essentially novices.

It is striking that the judgements made on the examination papers suggest that, especially at GCSE, the history papers were much more demanding than the geography papers. What tends to happen is that experts see little demand in the assessment of knowledge, seeing it rather in reasoning and evaluation requirements. This is in many ways a valid hierarchy, but what it ignores is that, for novices, simply acquiring and retaining the knowledge is much harder than experts recognise. In fact, the position is almost simpler than that. Faced with a very closed question, a candidate either has the specific knowledge to answer it, or not. It is therefore either an easy question or an impossible one. The way that the geography papers were constructed at GCSE meant that lower attaining candidates who may have had some real understanding of geographical principles, but little detailed knowledge, were simply found out.

What all this means is that, in terms of this study, it is very hard to draw conclusions about the relative standards of performance in the two subjects. It is certainly the case that, at GCSE and AS, the geography candidates did not display the same level of knowledge or skills as those given the same grade in history. However, reviewers were clear that it was not helpful to draw simplistic conclusions from the pattern of judgements. Indeed, there was no significant difference between the subjects at A2, when the candidates were assessed in very similar ways. What the findings actually suggest is that the subject communities – perhaps especially the geography community – should review their

approaches to assessment. It is possible that the history examinations do not make sufficient demands on candidates' actual knowledge; it seems certain that geography examinations do not provide candidates with sufficient opportunities to display their understanding, and even their knowledge.

4.3 Outcomes of study 1b: GCSE, AS and A level sciences

The materials used in the science study were somewhat different, with scripts drawn from a range of performance. Reviewers saw candidates who had achieved from mid-grade G to mid-grade D at foundation tier GCSE, from mid-grade C to mid-grade A at higher tier GCSE, and from mid-grade E to mid-grade B at AS and A2 (that is, 135–225 uniform marks on their AS units and 135–225 uniform marks on their A2 units).

This allowed the estimates of comparability to be made across the range of subject outcomes. It was also the case that the nature of the assessment was similar in all science examinations within a level. Any coursework tended to carry very similar weightings at a given level and, at GCSE at least, it was specified identically across all four subjects considered. The absence of coursework was not a significant factor for this study. To that extent the task for the science reviewers was relatively straightforward. This should not lead to any underestimation of how difficult the task remained. For example, the syllabus review noted that the identical coursework requirements may actually have made different demands in different subjects. In any case, it remained remarkably difficult to review significant volumes of evidence of the performance of two complex but different cognitive processes and decide which was the better. Indeed, it is arguable that it was impossible, because to do so begs the question ‘better at what?’ In this case, the answer should be better at some Platonic ideal of, in this case, 'scientificness', which, because it neither exists nor would it be likely that there is any real consensus about it, clearly underlines the limitations in any conclusions about this work.

The analysis of the patterns of judgements made by the reviewers revealed no substantial nor consistent differences in standard of performance between any subjects at any level. Any differences detected tended to be less than that reflected by the level of precision that awarders are capable of in award meetings looking at a single subject.

This was particularly true at GCSE. Here the differences (the method used in the work and in the analysis mean that there are almost bound to be differences) were very small. In fact, the average difference identified at foundation tier was just under 1.4 per cent, and at higher tier just over 1.4 per cent, and there are few subjects in which the consistency of marking and precision of grading could be guaranteed to that level of accuracy.

At A level the picture was slightly less consistent. At AS, the differences were again very small, with biology and chemistry almost indistinguishable, and performance in physics judged to be marginally less impressive. Here too, however, the scale of the differences was well within the inevitable margin of error in the examination system and, more

importantly, the reliability of the judgements arising from this exercise. At the lower grades in A2, the picture was again one of great consistency, with all three subjects almost perfectly aligned. At the higher grades, however, there was found to be a slight divergence, with biology candidates performing slightly less well to achieve their grades than the chemistry candidates, with the physics candidates in between.

4.4 Outcomes of study 2a: A level biology, psychology, sociology

Awarding bodies provided the complete examination work of candidates who had been awarded just a grade A and just a grade E overall. At AS, this meant candidates who had gained 240 or 120 uniform marks in total, with even performance across the units. For the A level candidates, the specification for the work was that it should comprise candidates who had gained 240 or 120 uniform marks in their A2 units, irrespective of the overall grades obtained. No coursework was included in the review.

There are important considerations to bear in mind when interpreting the patterns of judgements in this study. In the first place, the range of subjects covered is much wider than those covered in studies 1a and 1b. In particular, there are much more obvious differences between sociology and biology than between the sciences as a group or even history and geography. This was part of the design of the work, in that the focus was intended to be on psychology and the other subjects were chosen to reflect those aspects of the subject that make it a social science and those that make it more like a science. However, that design must make issues of bias much harder to predict and very hard to control.

The situation is further complicated by the fact that the A2 sociology scripts supplied for the study were from candidates whose results were somewhat better than those gained by other candidates. It is unclear what effect this might have on the kind of judgements that the reviewers might make.

The final factor to bear in mind for this study was also raised in the reflections on study 1a. The nature of the assessments for sociology and for biology were very different. This is made clear in section 3 of this report: the reviewers were clear that much of the demand in the biology examinations lay in the large volume and high cognitive level of the subject knowledge required. In this study, the question papers for biology were found to be relatively undemanding in the CRAS analysis. This accurately reflected differences in the approach to assessment, but the reviewers stressed that this did not fairly reflect the actual impact of the examinations on the candidates, because the factors used in the CRAS analysis did not address subject content. However, although the reviewers recognised this, it is hard to know what impact it had on their judgements.

At AS, all three subjects seemed very well aligned across the grade range. Any differences were well within the reliability of operational marking and, even more importantly, the confidence limits of this study.

At A2, there was some evidence that the performance of sociology candidates was not as impressive as that of candidates in either psychology or biology. The latter two subjects were very well aligned. Quite how much weight should be placed on the findings for sociology is hard to establish, especially in the light of the reservations expressed previously. What is perhaps more interesting is the fact that the analysis suggested that standards in biology and psychology were very well aligned across the grade range in both the AS and A2 examinations. Given that the initial impetus for this work was the suggestion that students were turning away from science to psychology because it was perceived to be the soft option, the study suggests that this perception has little basis in fact, at least in terms of the demand of the examinations and the grading standards set.

4.5 Outcomes of study 2b: A level English literature, history, media studies

Awarding bodies provided the complete examination work of candidates who had been awarded just a grade A and just a grade E overall. At AS, this meant candidates who had gained 240 or 120 uniform marks in total, with even performance across the units. For the A level candidates, the specification for the work was that it should comprise candidates who had gained 240 or 120 uniform marks in their A2 units, irrespective of the overall grades obtained. No coursework was included in the review.

As with all the other studies there are specific contextual factors that should be taken into account alongside the findings. Here, the approaches to external assessment were fairly similar across the three subjects, but media studies involved a significant technical dimension that was simply not present in the other studies. This technical dimension also meant that the proportions and nature of coursework differed more for media studies than for the other two subjects. As no coursework was included in the study, reviewers did not see media studies candidates' practical productions, worth 40 per cent of their overall A level and representing a very different set of skills. This made it harder for the reviewers to make confident judgements.

In addition, there were some aspects of the materials reviewed that may have affected the reliability of the outcomes. First, the A2 candidates in English were actually somewhat lower attaining as a group than those for the other two subjects. As with the similar issue for sociology in study 2a, this may have had an impact of the judgements made. Also, at AS, the range of work looked at for both English and history was very narrow.[13] This made the analyses difficult to quantify.

At AS there was very little difference between the subjects at the grade A boundary. At the grade E boundary the performance of candidates in media studies was considered to be slightly less secure than those in English, with history candidates in between. At A2, at both boundaries the media studies candidates were considered to be less impressive than the English candidates, with the history candidates in between.

[13] This is not a criticism of the awarding bodies for supplying the wrong materials. In fact, the scripts for English and history at AS complied most closely with what was specified. It was rather a flaw in the design of the study since a very narrow range of work made it hard to estimate the size of the differences found.

5 Conclusions

There are two types of conclusion about this work. Perhaps the most important is that the work is feasible. Although reviewers found it challenging, none suggested that it was impossible or that any findings would be invalid. What is more, the project was able to make use of and build on a taxonomy of cognitive demand that has real potential as a metric for qualifications.

The work also highlighted the complex nature of issues surrounding comparability. It was evident that measures taken to help ensure comparability, such as the use of common assessment objectives in the sciences, may have a variable and unforeseen impact in different subjects, and do not necessarily lead to comparability.

In terms of findings, overall, the key one was that subjects were generally in line. In particular there was little evidence that the A level subjects sometimes described as ‘soft’ were any less demanding than their more established counterparts.

The one area where a significant difference in performance was identified, between geography and history in study 1a, highlighted differences in approach to assessment rather than standard. (This was replicated to an extent when biology, psychology and sociology were compared, where the cognitive demand of the question papers in biology was lower than that for psychology or sociology, and the real demand of the subject lay in the need to recall large volumes of complex content.) This is a very useful finding. While it highlights a weakness in the approach, it serves to focus attention on validity: the differences in approach to assessment can only really be justified by a thorough review of the aims and objectives of the subject.

