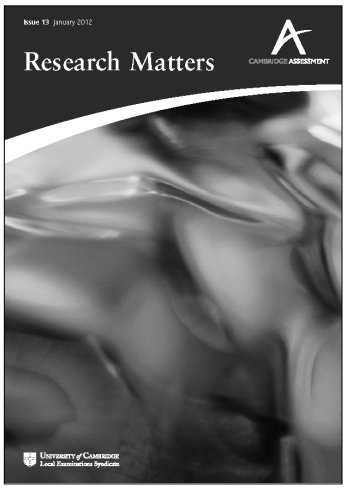
Issue 13 January 2012

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate



If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters : 13

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

Innovation, creativity, learning to learn, critical thinking, 21st Century Skills. These are mentioned in many of the policy statements on education and training, in many nations – not just England. One of the contributions that an assessment organisation can make to the discussions of these ideas is that of clarity. Not pedantic over-analysis, but clarity. Validity in assessment is vital. We need to understand what we are assessing and design our assessments to apprehend it in a reliable way.

It is therefore sensible to ask whether something such as 'creativity' exists and in what way does it exist? Some individuals appear more creative than others. In what way do they vary from people who appear not to be creative? Is creativity limited to certain areas of human endeavour, or can it exist in any subject discipline? If you are creative in one context are there limits on whether it can be 'transferred' to other contexts? These are only the opening set of questions. And they count, since in most assessments we are making a claim about the skills, knowledge and understanding of individuals and, in many cases, using the assessment to make inferences around the future performance of an individual.

As with others listed above, 'Critical Thinking' is a domain which remains in need of clarification, both in respect of learning and assessment. There is a growing international literature into which the Research Division has tapped, and results so far suggest that deliberately focussing on critical thinking as part of curriculum planning, delivery and assessment can elevate attainment. From this we are beginning to understand the approaches which work. But there is evidence of approaches which do not. As a result of this discrepancy, Beth Black has undertaken to clarify and better structure 'Critical Thinking' through her editorship of the *A to Z of Critical Thinking*, which we hope will be a major contribution to the field.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

The themes addressed in this issue reflect the diversity of research carried out at Cambridge Assessment.

In the first article Shaw discusses an assessment approach where a curricular subject is taught through the medium of language other than that which is normally used and considers the linguistic, educational and cognitive challenges across a number of subjects. His research has already informed practice and his proposals for future work in this area highlight future needs.

The work from Vidal Rodeiro on special consideration enhancements was presented at the annual British Educational Research Association (BERA) conference in September. This focussed on patterns of special consideration applications over time, for different qualifications, by school type and by outcomes. Little research has been conducted in this field and Vidal Rodeiro's work is a welcome addition to the literature.

Considerable research has been carried out at Cambridge Assessment over many years on the challenges of judging the quality of scripts. This is a fundamental part of our assessment process and Bramley adds to the debate by reporting on the features of examinees' scripts that influence judgements of quality. Although Bramley identified some problems with the method, his approach provided a new way of investigating a difficult problem and he suggests further research to improve validity in this area.

Suto and Nádas discuss the importance of research and project work for 16 to 19 year olds, outlining the diversity among research routes and the breadth of skills that are enhanced through such study. They also highlight some of the challenges inherent in assessing such achievements and identify important curricular and assessment issues that need to be considered as qualifications for the future are developed.

Haigh also presented his work on item design in computer-based assessments at the BERA conference. He highlights the importance of fairness for students undertaking assessments and the need for us to be aware of any unintended consequences of moving from paper-based to computer-based testing.

The final article reports on Cambridge Assessment's Information Services Platform and the innovative strategy it represents. Raikes explains how the platform development allows us to harness the data we now have available to enhance quality assurance processes.

Sylvia Green *Director of Research*

International assessment through the medium of English: analysing the language skills required

Stuart Shaw CIE

Introduction

International assessments in a wide range of subjects are being prepared for and delivered through the medium of English. These are taken by many candidates whose first language is not English and increasingly by students who have participated in Content and Language Integrated Learning (CLIL) programmes in a range of different linguistic contexts. CLIL – defined as "an approach in which a foreign language is used as a tool in the learning of a non-language subject in which both language and the subject have a joint role" (Marsh in Coyle, 2006, p.1), involves "learning to use language and using language to learn" (Marsh and Lange, 2000). The CLIL approach consists of teaching a curricular subject through the medium of a language other than that which is normally used and operates in a range of contexts and is subject to varying interpretations. In CLIL programmes of learning, learners gain knowledge of the curriculum subject while simultaneously learning and using the foreign language: curricular content leads language learning. Interaction in learning – a fundamental tenet of CLIL, is important because learners need to use and develop language *of* learning (the content); language *for* learning (peer interaction); and language *through* learning (for cognitive skills).

Typically students preparing for University of Cambridge International Examinations ('Cambridge') do so in very diverse linguistic and educational contexts, some following an entire curriculum in English, and others undertaking only one or two Cambridge examinations in parallel with qualifications from their own (non-English) national curriculum. The integration of two curricula in bilingual education programmes presents challenges for the schools and interesting issues for Cambridge. Cambridge is keen, therefore, to understand this context in order to evaluate the impact of this choice of education programme and particularly the role of assessment within it.

The Cambridge context raises a number of issues relating to *language awareness* (e.g. progression from basic interpersonal communication skills to cognitive academic language proficiency) and *assessment* (e.g. the level of English needed to access and succeed in international assessments).

The focus for the study described here is the International General Certificate of Secondary Education (IGCSE). The study adopts a two-phase methodology and involves an analysis of language use in Geography, History and Biology in order to (a) build a 'profile' of the language skills required and evidenced by IGCSEs and (b) determine whether any identifiable linguistic patterns adhere to different content, non-language IGCSEs.

Language, educational and cognitive development

Students studying content subjects in a second language (L2) need to demonstrate competence not only in their familial linguistic background (L1) but also within the educational community in which they are required to function or operate. This raises issues relating to bilingualism. Although it is not the purpose of this article to explore definitions of bilingualism, bilingualism is used here to refer to the use of two or more languages to operate in society, without regard to the level attained (see Mackey's [1968] use-based definition). Grosjean (1982, p.220) – with reference to the earlier work of Jespersen (1922), points out a child

> *hardly learns either of the two languages as perfectly as he would have done if he had limited himself to one*.

Language acquisition has clear implications, therefore, for a learner's educational development:

> *The brain effort required to master two languages instead of one certainly diminishes the child's power of learning other things which might and ought to be learnt*. (Jespersen 1922, in Grosjean, 1982, p.220)

Language development

A number of linguistic idiosyncracies have been observed amongst students who exhibit language competence in two or more languages, particularly amongst bilinguals (Kelley, 1936; Tireman, 1955; MacLaughlin, 1978). These include limited vocabularies and grammatical structures, unusual word order, errors in morphology, hesitations and stuttering. MacLaughlin (1978) has argued that such difficulties are less to do with the process of bilingualism but more the fact that such children are forced to learn a second language in the school and do not have equal exposure to the two languages. An alternative thesis is offered by Skutnabb-Kangas and Toukomaa (1976) who have proposed the *Developmental Independence Hypothesis*. This hypothesis suggests that if the L1 is poorly developed, then focus on the L2 will impede the continued development of the L1. As a consequence, development of the L2 will be inhibited and lead to 'semi-lingualism'.

Cummins (1976) has suggested that children can – contingent upon teacher, home and community support, become bilingual at no cost to their L1. In his *Common Underlying Proficiency Theory* (Cummins, 1980), Cummins argues that the two languages used by an individual, though on the surface apparently quite separate and distinct, function through the same central cognitive system or as Baker asserts: "When a person owns two or more languages, there is one integrated source of thought" (Baker, 1996, p.147).

Educational development

Some research indicates that learners who have been required to develop linguistic competence in two (or more) languages lack both interest and initiative and have, as a consequence, fallen behind educationally (Macnamara, 1966; MacLaughlin, 1978). According to MacLaughlin (1978), any educational impediment can be accounted for by testing

content in a second language over which the child has not yet developed sufficient command, combined with other minority ethnic factors such as low socio-economic status and negative attitudes of the majority group. In order to address the problem of insufficient command of L2, Cummins has proposed the Basic Interpersonal Communication Skills (BICS)/Cognitive Academic Language Proficiency (CALP) distinction.

The acronyms BICS and CALP are commonly used to discuss the language proficiency levels of students who are in the process of acquiring a new language. In an attempt to understand progression in students' learning of content and language, Cummins has shown how students need to progress from BICS (low cognitive demand, context embedded) towards CALP (high cognitive demand, context reduced). The distinction was intended to highlight the different time periods experienced by students to acquire conversational fluency in their L2 as compared to academic proficiency in that language. CALP is a language-related term which refers to formal academic learning, as opposed to BICS which are language skills needed in social situations. Typically, students develop proficiency in BICS well before they acquire a strong grasp of CALP: conversational fluency is often acquired to a functional level within about two years of initial exposure to the second language whereas development of academic aspects of the second language often takes between five and seven years (Cummins, 1981; Collier, 1987; Klesmer, 1994). As a consequence, students may give the appearance of being fully proficient and fluent, while still struggling with significant academic language deficiencies.

From a pedagogic perspective, the BICS/CALP distinction helps teachers support students to access cognitively challenging content material by embedding activities in a supportive context.

However, the BICS/CALP distinction is not without its detractors:

- the distinction reflects an autonomous perspective on language that ignores its position within social practices/power relations (Edelsky *et al*., 1983; Wiley, 1996).
- CALP promotes a 'deficit theory' in that it attributes the academic failure of bilingual/minority students to low cognitive/academic proficiency as opposed to inappropriate education (Edelsky, 1990; Edelsky *et al*., 1983; Martin-Jones and Romaine, 1986).
- CALP represents little more than 'test-wiseness' (Edelsky *et al*., 1983).

The BICS/CALP distinction continues to engender debate. Notwithstanding the arguments, the distinction has had a longstanding effect on education and bilingual education in particular and is promulgated in strategic policy. For example, Tucker (1999 website) comments that the study carried out for the World Bank by Dutcher in 1994 concluded that:

> *the best predictor of cognitive/academic language development in a second language is the level of development of cognitive/academic language proficiency in the first language* and that *cognitive/ academic language skills, once developed, and content-subject material, once acquired, transfer readily from one language to another.*

Related to BICS/CALP are the concepts of *content-obligatory* and *content-compatible* language. When learning content through a second language, it is a requirement for a student to produce both content-obligatory and content-compatible language in a potentially wide range of subjects.

Content-obligatory language or specialist language is the language that can be taught in the context of a particular subject and is essential to an understanding of content material. This is the subject-specific vocabulary, grammatical structures and functional expressions learners need in order to be able to learn about a curricular subject, communicate subject knowledge, and participate in interactive classroom tasks. In the context of History, for example, learners can discuss history either using general historical terms and phrases that are needed to operate within the subject but are not tied to a given period (e.g. collapse, defeat, democratic), or using words and phrases relating to the specific periods/events studied, which mainly amounts to nouns and proper nouns (e.g. conscription, hyperinflation, treaty).

Content-compatible language is language that can be taught naturally within the context of a particular subject matter and that students require additional practice with. This is non-subject-specific language which learners may have been exposed to and learned in their English language classes and which they can use in CLIL classrooms to communicate more substantively in the subject.

Examples of content-obligatory and content-compatible language in the context of Biology are shown in Appendix A.

Cognitive development

The literature on the cognitive effects of language learning is mixed. Some research suggests that foreign language education increases cognitive development and positively influences academic achievement in other subjects. Stewart (2005) cites previous studies that found positive correlations between bilingualism and non-verbal measures of cognitive ability in young children. Grosjean (1982) notes that whilst some research indicates no effect on cognitive growth (Barik and Swain, 1976), other researchers have claimed negative effects (see Saer, 1926; Darcy, 1946). Lambert (1977) argues, however, that early IQ studies were beset with research methodology weaknesses (including not controlling for age, sex, socioeconomic background, educational opportunities, degree of bilingualism, matching on too few factors, lack of test adaptation for the linguistic minority).

Peal and Lambert in 1962 claimed French-English balanced bilinguals to be superior intellectually – scoring higher on both verbal and non-verbal IQ tests. However, the authors did concede that it is not clear whether intelligence is the reason for such an outcome. Others have also argued that bilinguals can have superior thinking abilities based on their dual linguistic systems. Garcia (2009) cites Vygotsky (1932) who contended that bilingual children had two ways to describe the world and so had more flexible interpretations. Garcia also notes work by Scott (1973) who reported more divergent thinkers amongst bilinguals when he told subjects to think of an object and say as many things as possible that they could do with it. Garcia notes that such studies show that bilingual children tend to give more responses, which are original and elaborate.

In attempting to resolve the conflict between the positive and negative effects, Cummins (1976) has suggested that there may be a threshold level of linguistic competence which a bilingual child must attain both in order to avoid cognitive deficits and allow the potentially beneficial aspects of developing bilingualism to influence their cognitive functioning. The 'Threshold' Theory was first put forward by Toukomaa and Skutnabb-Kangas in 1977. It suggested that the development of two or more languages in a balanced bilingual person moves upward through three identifiable levels, crossing two distinct thresholds in between levels. According to this theory, positive cognitive advantages are only to be achieved when the first and second thresholds have been crossed.

The International General Certificate of Secondary Education (IGCSE)

The focus of this study is the International General Certificate of Secondary Education (IGCSE).[1] The IGCSE is taken in a range of subjects at the end of a two-year course. At a similar and recognised level to the UK General Certificate of Secondary Education (GCSE), the IGCSE was developed for a global market, striving for non-UK centric contexts and awareness of second language needs. The IGCSE is open to schools from all over the world and is available twice a year in June and November. In many subjects there is an Extended and a Core Curriculum. The Extended Curriculum includes the material from the Core Curriculum, as well as additional, more advanced material. Each learner's performance is benchmarked using eight internationally recognised grades: Extended Curriculum: A*, A, B, C, D, E; Core Curriculum: C, D, E, F, G.

Research questions

This study sought to address the following questions:

- What level of English, according to the Common European Framework of References for Languages (CEFR), is needed to access and achieve in typical IGCSE assessments?
- What cognitive and academic language skills are needed to access and succeed at typical IGCSE assessments?

Key specific linguistic questions for both phases of the study were organised under three principal themes:

Lexical, structural and functional resources

- What are the main language functions that students are being asked/demonstrating in their answers?
- Is there a pattern in the occurrence of structural forms of a particular type?
- Are examples of assessment specific vocabulary clearly comprehensible from syllabus guidelines?
- What are examples of subject specific vocabulary and what proportion of test questions mention or require responses involving subject specific vocabulary?
- Have candidates understood assessment specific vocabulary and effectively applied the requirements appropriately in their responses?

Expected and actual candidate performance

- What writing skills required in mark schemes were anticipated/reflected in candidate responses?
- To what extent were candidates penalised by the ineffective use of subject specific vocabulary?
- What are the observations of the Principal Examiner on the use of language?
- How does candidate use of language compare with analysis of question papers and mark schemes?

Criterial task features relating to student performance

- Is there evidence of undue cognitive reading demand made of candidates?
- What is the typical length, format and complexity of question input and rubrics/candidate responses?
- Is achievement linked to length of response?

Methodology

The first phase of the study focused on Geography, History and Biology from the November 2008 and June 2009 sessions and entailed an analysis of syllabuses, question papers and mark schemes to allow a full overview of the qualification. In addition to analysing the June 2010 question papers and mark schemes, Phase 2 also involved an analysis of candidate performances, consultant and examiner reports.

Focus was on the written components (as opposed to practical or coursework components). In order to obtain varied perspectives on each IGCSE, four grade levels were sampled (A, C, E and F) from four linguistic backgrounds (Romance; Semitic; Sinitic; Slavic). As Biology includes a multiple choice paper and requires shorter written responses in candidate scripts, five candidates at each grade were studied, whereas in History and Geography three candidates were sampled at each grade.

The final data set comprised 74 Biology scripts; 47 History scripts; and 48 Geography scripts. Additional documentation was provided for the second phase of analysis in the form of reports on the issues of language written by senior examiners. These reports together with Principal Examiner insights enabled Phase 2 to be located in a broader context. (Principal Examiners are responsible for standards in the setting of question papers and the marking of examination scripts.)

Findings

The findings are presented in terms of:

- the minimum level of English competence required to access and succeed at IGCSE;
- how the linguistic demands in the qualification might relate to the CEFR;
- the extent to which the language competence demonstrated could be defined as CALP.

In order to understand how the findings relate to the CEFR, a short description of the purpose and structure of the CEFR is provided.

Designed as a guideline to describe achievements of learners of foreign languages across Europe (and increasingly in other countries), the CEFR is a framework that provides a basis for the mutual recognition of language qualifications and enables awarding bodies to define and articulate language proficiency levels and interpret language qualifications.

The framework identifies six levels of potential language proficiency, two at basic language user level, namely A1 *Breakthrough* and A2 *Waystage*; two at independent user level: B1 *Threshold* and B2 *Vantage*, and two at proficient user level: C1 *Effective Operational Proficiency* and C2: *Mastery* level. The six reference levels are becoming widely accepted as the European standard for grading an individual's language proficiency. To illustrate these levels, CEFR global scale reference level descriptors

1. http://www.cie.org.uk/qualifications/academic/middlesec/igcse/overview

have been provided as Appendix B. The reference descriptors constitute a superordinate set of specifications, among nearly 60 scales provided by the CEFR (Council of Europe, 2001) to define different language skills, communicative purposes, contexts, activities, modes, etc.

The CEFR scales are intended to inform the development of language curricula, courses, tests and other forms of assessment, summative and formative, external and internal. The CEFR has growing relevance for language testers and examination boards, helping to define language proficiency levels and interpret language qualifications.

Specific findings from each of the two phases of the study are now reported.

IGCSE History

The input language used in IGCSE History is of a high level. The language of the rubric falls mainly within the B2 level of the CEFR in terms of structural and lexical load. Although the rubrics and questions are generally expressed clearly using accessible language and could be understood by a B2 level student, the lexical input of the accompanying stimulus material is much higher and students would need to be at least C1 level to be able to process the text. There are many examples of structurally complex input including cleft sentences; organisation in terms of desired thematic prominence (rather than for accessibility or simplicity of structure); reported speech using a range of verb tenses, relative clauses and conditional structures.

Candidates need to be able to cope with a significant amount of subject-specific language, meaning that CALP is required. The question papers, and the source material which the questions refer to, contain a large amount of subject-specific vocabulary. Generally, this vocabulary falls into two lexical categories:

- **general historical terms and phrases** needed to operate within subject but are not tied to a given period (e.g. collapse, defeat, democratic and phrasal verbs such as set up, step in, take away)
- **specific lexis** that is linked to certain periods or topics (usually nouns and proper nouns such as conscription, colony, hyperinflation, dissidents, treaty)

Source texts may contain low frequency language and be challenging in terms of their 'authenticity'. Some of this material is complex (for example, the fronting of sentences with complex noun phrases) placing a high cognitive reading demand on candidates who are expected to quote from the material, and use it selectively in the exemplification of their argument. It is envisaged that candidates are prepared for this fact, and will also have in-depth knowledge of the historical period in question.

The use of cartoons and artwork in the input may pose challenges in terms of cultural non-familiarity though they may help to lessen the reading load. Their selection engenders interesting issues of accessibility and cultural relevance, and their appearance on papers may cause different challenges for candidates in different parts of the world.

Although candidates do not need to read source texts much longer than 250 words, they do, however, need to demonstrate a range of reading skills and strategies including careful reading at global level; careful reading at clause / sentence level; intensive reading of data; dealing with unfamiliar words and referencing skills (including exophoric referencing to link what they have read to a wider historical context).

The use of high-level input information used to set the scene for History questions suggests that emphasis is being placed on the top-down processing model of language or reading comprehension. This is a model based on the belief that readers make sense of discourse by moving from the highest units of analysis to the lowest, and that comprehension is achieved by firstly activating background knowledge or schemata and setting the context.

Questions range from those requiring short answers (low tariff) to those requiring longer answers (high tariff), which ask for opinion, evaluation, justification and explanation with reference to source material. Short answers can be written at word or phrase/sentence level, but more open questions require longer, coherent answers usually consisting of more than one paragraph.

If the exemplification in the mark scheme for Paper 1 (consisting of questions selected from the 19th century and 20th century 'Core' topics) were to be seen as typical of the target output, candidates would be expected to produce language that is well above B2 level of the CEFR, even if the content-specific language is disregarded.

In terms of their written ability, IGCSE History candidates need to be able to demonstrate a range of writing skills. Students learning about History are required to be able to organise their ideas clearly, in order to present effective and balanced arguments that show evidence of evaluation and interpretation. They also need to demonstrate concision in certain questions and extended reasoning in others. They need to be able to quote judiciously from sources and exemplify claims from their knowledge. History teachers may need to teach this language or at least make learners aware of it in order for learners to be able to use it effectively

While accuracy of surface features such as spelling, word order, and grammar may not be fully mastered, candidates need to have a solid repertoire of structures, together with a wide vocabulary range. This will include many subject-specific terms and a number of nouns and proper nouns relating to the specific periods they have studied.

An important observation from the second phase is that low marks usually stem from deficiencies in the subject – lack of recall, inaccurate claims, unsupported assertions, one-sidedness, misinterpretation of question or sources, failure to evaluate, etc – rather than any obvious linguistic shortcomings.

IGCSE Biology

The input language used in IGCSE Biology is not of a very high level. A student with B2 level English could do well on this qualification. There seems to be little or no requirement for detailed explanations or reasoned speculations, both of which would require an advanced level of English. Whilst students with B1 level English could access the paper and understand the questions they would not have a flexible enough command of English at their disposal to allow them to work at the speed required to complete the paper in the time given. Knowing the answer is the first step but having the language resource to describe processes/factors/differences with limited drafting time is a B2 level skill.

Generally, rubrics and questions are clearly written and a simple sentence structure is used (usually employing imperatives). The structures are often repeated. It is rare for any one part of a question to take up more than two lines and the layout is spacious and accessible.

Candidates have to read and understand a range of forms of input: graphical data (diagrams, tables), photographs, short/long questions, instructions (for the practical test). As with History, candidates need to employ a range of reading skills.

A number of different functional verbs are used for Biology, and each has a precise use and meaning though these subtle distinctions may not

be clearly understood by teachers and candidates who will have encountered these verbs previously in different language learning contexts. There is, therefore, an added level of challenge required to recognise the exact force of these verbs and produce what is required. Consequently, there is a far broader range of language functions involved in Biology than in, say, History and the subtleties underpinning the different verbs will have to be mastered, if candidates are to succeed.

The wording of questions is kept simple and structural forms are controlled within this. Gap-filling tasks are a good example of structural simplicity, and would be fully accessible to students from B2 level. Some question types involve processing and deduction. However, factual points are made clearly in single sentences which are then separated by a line space to assist candidates in their reading. This type of question is balanced by others with minimal text and which include the visual support of diagrams or illustrations.

On the multiple choice question Paper 1 (consisting of four-option multiple choice where candidates have 45 minutes to work through forty questions of differing formats, some including illustrations and others text and tabulated data), candidates clearly have to work at speed, reading efficiently. The reading load is not excessive on this paper though certain questions involve four full-sentence options.

The level of content specific vocabulary is very high across the papers. The learning of subject-specific terms for Biology is inextricably linked with the learning of the subject itself, in a way that is very different from, say, the learning of History.

Candidates often have the visual support of diagrams for a science subject (the level of graphical data input is high with a majority of the questions comprising graphical data in some form). However, there is inevitably a huge learning load, and all questions use subject-specific vocabulary, even if the responses do not always require it.

Language competence does not impact on Biology as much as in History. The Biology student is not required to produce long developed/reasoned answers: the mark scheme does not award marks for reasoning and development. Most of the answers requiring continuous prose are descriptions which can be done successfully with simple structures and key content-specific terminology.

Phase 2 reflects the findings of Phase 1: the language used in the Biology question papers is generally quite simple, with predictable structural forms and a limited range of command terms. However, the subject-specific vocabulary is much more demanding and key terms can be found in almost every question in each paper. Candidates who do not have a good grasp of this vocabulary would struggle to complete the questions.

Candidates are not required to produce long, detailed pieces of writing and many answers can consist of single words, short phrases or a few sentences at most. Where longer sentences are produced, most can be written using present simple or present continuous tenses, active and passive forms, basic conditional structures, comparatives or imperatives. Whilst candidates need to be able to produce these structures, conveying meaning appears to be more important than accuracy of expression.

Marks are only awarded for stating facts or identifying factors, reasons and so on – the style in which the answer is written does not matter. However, as well as naming things, stating facts or defining terms, candidates are required to interpret information and data, speculate, make suggestions and give detailed explanations, all of which are academic skills which need to be learnt in the classroom. Topics may not be immediately familiar to candidates and they may be required to make connections to the subject matter they have learnt, draw conclusions or apply their scientific knowledge to a given situation. Therefore, CALP is important to some degree – candidates need to be able to assimilate information, know what type of information is required for each question, be able to make links, apply knowledge, and so on. Those candidates who give descriptions of what they see in a diagram rather than interpreting it would not be awarded marks; those who only state a fact but ignore the instruction to also give an explanation would be heavily penalised.

IGCSE Geography

The input language used in IGCSE Geography is not of a very high level. Generally, a B2 level student would be able to cope with the vast majority of the rubrics, questions and input material. In the question papers, assessment-specific vocabulary appears in the rubrics and in the questions themselves, giving instructions and specifying the functional language which candidates are required to produce. Like History, candidates are required to identify from the rubrics the functional language required.

Candidates have to read and understand a range of forms which include graphical data (diagrams, bar charts, pie charts, maps, tables), photographs, short/long questions, short texts. The volume of graphical data is high but much of it can be understood only if the accompanying text is understood. In all three papers candidates are required to scan input material (whether it is a table, map or text extract) to locate answers. Candidates are also required to read intensively for detail. This entails reading a wide range of graphical data carefully; separating data from questions; reading numerical and other information from graphical data accurately; and moving between graphical data and text.

The papers contain a mixture of closed and open question types, requiring answers of varying length and format though the length of the answers is not specified. Overall, there is not a significant amount of extended text for candidates to read in any of the question papers, however all the papers consist of several questions, which each have a different number of sections and sub-sections. As a consequence candidates need to employ a variety of reading skills.

In Paper 1 (in which questions are resource based, involving problem solving and free response writing) they need to skim read the six questions in order to choose which three questions to answer. This involves reading the whole question with all its sections to check which information on which aspects of the topic is required for each section. Candidates must ensure that each section is answered and repetition/overlap of information is avoided. Candidates also need to read stimulus texts through before answering.

As papers do not have a standard format, candidates need to concentrate to read different question formats and different question types. Candidates may also need to deal with unfamiliar lexis which would entail deciding whether the unknown word is a key word and determining linguistic clues (using pictures/diagrams). Candidates need to be able to read the rubrics and questions carefully at clause and sentence level in order to be able to identify the type of response required (key words in the rubric) and what functional language to use in their answers. This can sometimes involve sophisticated recognition of textual patterns.

Like History and Biology, candidates are required to be able to cope with content-obligatory language with most questions containing some subject-specific vocabulary. Some of this vocabulary is not very high

frequency and may be challenging at this level. Whilst the level of content specific vocabulary is quite high – with some questions comprising higher-frequency language than others – all questions require candidates to understand subject-specific vocabulary and then to produce appropriate subject-related vocabulary in their answers.

There is an expectation that the candidate has flexible language resources to deal with a wide variety of question types. A B2 level student should be able to produce adequate responses, providing they have the lexical range, but CALP is required. Short answers require quite specific content-based language; longer answers need content knowledge but also a range of language to be able to describe, explain and draw conclusions, as well as the ability to write concisely.

Geography is a subject where students not only have to learn how to work with data, but also how to communicate in writing a wide range of concepts and ideas. Students with a good knowledge of Geography learnt in L1 would struggle to 'translate' this knowledge into English unless they had advanced language skills. Students who study Geography in the context of the English language would have a huge advantage when coming to these papers where language competence plays a key role. They would have learnt the subject while also learning to explain why things happen or might happen in English. C1 level students with a good knowledge of Geography and good data skills would perform well on this assessment. They would be able to write concisely for short answers, reformulate and develop ideas and speculate in longer answers, drawing on ideas learnt during the course as well as on evidence in the data on the paper. They would have language resources such that they could construct cohesive and coherent answers at the speed required (the mark scheme rewards development when longer answers are called for).

Analysis of scripts reveals that all candidates are able to attempt the majority of the questions. The main issues in terms of language use are (i) format of answers, that is, note form, bulleted lists, longer explanations; (ii) the range of language used; and (iii) the accuracy of the language used.

Candidates were able to use a good range of subject-specific vocabulary. Some of the vocabulary has a more general meaning but is relevant to and appropriate for the topics in the papers. In addition to using subject-specific vocabulary, candidates also demonstrated successful use of a range of general language structures and expressions.

Two issues of interest with regard to candidate performance are, first, the ability to produce developed answers and, linked to this, the ability to deal with questions requiring some form of speculation and judgement. To quote the comment from the Principal Examiner on the June 2010 Paper 4 (Alternative to Coursework which includes questions involving an appreciation of a range of techniques used in fieldwork studies):

> *Weaker candidates scored on 'practical questions, such as drawing graphs' while candidates 'of higher ability' scored well on the 'more challenging sections requiring explanation and judgement, especially hypotheses'.*

Discussion and conclusions

IGCSE alignment to the CEFR

On the evidence of this study, candidates entering for the IGCSE History examination will be above B2 level and those attaining A and C grades will be at C1 or above.

Although many of the IGCSE Biology candidates are of a very high level and may even be bilingual, a minimum level of B2 on the CEFR is required. This is in part due to the high level of subject-specific language that they are required to cope with, but also because not all the topics are immediately familiar to candidates, the fact that some evaluation or synthesis of information is required, and that key points in explanations need to be made clearly and without ambiguity.

For IGCSE Geography, the level of output of grade A candidates is certainly C1 in terms of range, accuracy and control of collocation. Candidates scoring lower grades are writing at B2 level and sometimes below. Although the approach to accuracy is not explicit in the mark schemes, it is assumed that comprehensibility of the answer is crucial as there is evidence that answers with non-impeding errors and only very basic cohesion score marks for content. In terms of the questions requiring explanation, speculation and judgement, the level of language in successful answers is closer to C1 than B2. Explaining content in black and white terms can be done at B1/B2 level but to qualify ideas, to describe the colours in between – is an advanced language skill.

Therefore, it can be concluded that a minimum CEFR level of B2 is useful to access typical IGCSE subjects, and that a CEFR level of C1 could provide an added advantage of linguistic resources to be able to develop arguments needed for higher grades for Humanities subjects such as History and Geography. Each subject necessarily requires different types of CALP.

	IGCSE		
	History	*Geography*	*Biology*
Overall CEFR alignment	B2/C1	B2/C1	min B2
CEFR User level	Independent/ Proficient	Independent/ Proficient	Independent
Requirement for CALP	✓	✓	✓

Supporting language claims underpinning the IGCSE

IGCSE claims an international reach and a local relevance as illustrated in the following quote taken from a current IGCSE Handbook:

> *The syllabuses use international examples and avoid terminology only used in one country. Non-native speakers of English are always treated fairly.* (Cambridge IGCSE, 2010, p.11)

Fairness is concerned with "the consequences of testing for individuals, groups or society as a whole" (Davies *et al*. 1999, p.199) and is a social rather than a psychometric concept. Because fairness has no single meaning there is, therefore, no single definition. The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME) note four possible meanings of fairness: (1) as requiring equal group outcomes; (2) as a lack of predictive bias; (3) as requiring that candidates have a comparable opportunity to learn the subject matter covered by the test; and (4) in terms of the equitable treatment of all candidates. One aspect of equitable treatment relates to the provision of reasonable accommodation for test takers with linguistic 'disadvantages'. To what extent then do the findings from this study substantiate the claim of fairness in terms of equitable treatment of all candidates?

The reading load in Biology is not high and it would seem that the quality of candidate responses depends less on time pressures than on

the ability to clearly express the required information. Thus marks are awarded according to the information stated rather than how it is expressed. Many of the questions across the Biology papers call for single words, short phrases or short descriptions. Therefore, candidates whose command of English is not fluent should still be able to complete the papers. The mark schemes do not describe specific writing skills that are required of candidates. However, it can be inferred from the mark schemes that single words and short phrases are acceptable, and that candidates do not gain extra marks by constructing complete sentences or longer, more coherent paragraphs. In general, there are few problems with candidates misunderstanding assessment-specific vocabulary – indeed many of them are using subject-specific vocabulary very effectively.

Whilst the language used in History is of a much higher level than Biology, examiners need to be congratulated for their tolerance of less than perfect English and their diligent processing of answers that are often dense, unparagraphed and written in challenging handwriting. The approach to marking appears to be positive rather than punitive, and any evidence, however thin, is likely to be sought out in order to raise an answer to the appropriate level of the mark scheme. Many candidates seem to perform effectively in English, which is a foreign language for them, apart from those few students who are fully bilingual. These candidates appear to be given every consideration both in terms of the questions they must answer and assessment of their responses.

Generally, there seem to be few problems with Geography candidates not understanding assessment specific vocabulary and most are able to provide answers appropriate to the question. Most candidates across the grades are able to complete all the questions and invariably with full answers. Even weaker candidates who score zero for many of their answers are able to write something for each question (sometimes at length and often with much irrelevance). In Paper 2 (based on testing the interpretation and analysis of geographical information and on the application of graphical and other techniques) and Paper 4 (the 'Alternative to Coursework' paper), many candidates, mostly those with a lower level of language range and accuracy, answered questions with phrases and bulleted lists, often with fractured grammatical structures. Stronger candidates produced full sentences and short paragraphs, almost always filling the lines provided for the response. Whilst the mark schemes make no reference as to whether both approaches are acceptable the assumption is that it is the content that counts, and not the style of the answer.

Interestingly, it is clear that candidates who use bulleted lists but have the linguistic resources to write full answers are penalising themselves unnecessarily. Those whose linguistic resources are not sufficient to support fuller answers can score satisfactorily on short-answer items (assuming subject knowledge) but cannot achieve maximum marks on questions requiring developed answers (and which often have higher totals of marks available).

In terms of relative time allowance, it is assumed that stronger candidates can produce longer and more cohesive text in the time given than weaker candidates. Sometimes, however, there is evidence of possible time advantage to candidates with knowledge of the correct answer and who opt for note form. In this case there is no evidence that providing lines for the answer guides candidates as to length; writing concisely is, however, a skill not always easy to acquire when writing in any language.

Research informing practice

It is hoped that findings from this research will help to raise 'second language awareness' in all stages of development of question papers, mark schemes and examiner reports. Findings have already contributed to the question writing process: question setters need to be aware of potential language issues confronting an international candidature.

Outcomes will also inform the construction of a 'CALP guide' – *Language Awareness in Teaching: A Toolkit for Content and Language Teachers* (Chadwick, *in press*) – designed (a) for teachers of content subjects who teach to students for whom English is not their first language; (b) for English as a Second Language (E2L) teachers who teach students who take some of their content subjects in English in other departments of their school; and (c) for content teachers who teach students for whom English is their first language. (English may be the teacher's first or second language but in this case we can assume their proficiency in English.)

The function of the toolkit will be:

- to provide content teachers with a place to find the kind of language their students need support with when studying for their IGCSEs, and language that will enable their students to engage with the content subject more effectively. This language will be CALP that is useful for all academic subjects and examinations;
- to help content teachers become 'language aware';
- to include a rationale and strategies for supporting students with this language in the classroom;
- to provide guidance to E2L teachers on how they can support content teachers and students taking content subjects in English in their school;
- to provide E2L teachers with a resource that they can use to help plan and supplement their English lessons to be more effective across the curriculum.

Future research

Building on the research reported here future studies will attempt to assess the impact of linguistic complexity and language accessibility on candidates taking international A level examinations designed for 16–18 year olds. The research is designed to comprise three phases. In phase 1, the marks obtained by each student for each sub-question on the exam papers for a random sample of at least 200 scripts for A level Geography and A level Physics will be collected and keyed into data spreadsheets. The data sets will be used to conduct a number of statistical analyses to describe question functioning for both whole questions and question parts using traditional and item response analyses. In phase 2, questions that statistical analyses suggest are performing in 'unexpected' ways (extremes of difficulty; reverse thresholds, a number of overfitting and underfitting items) will be explored further using textual and discourse analytic techniques in order to determine whether the questions present problems for international candidates and, more importantly, why these questions might be problematic. In the final phase of the research, students studying in their second year of A level Geography and A level Physics from a range of linguistic backgrounds will be asked to engage with the input language of questions identified in phase 1 and to comment on their linguistic complexity. Triangulation of textual analysis and think-aloud protocols will provide a powerful means to explore

complex syntactic and lexical features that challenge English language learners. Through the 'voices' of students, this work will scrutinise the appropriateness of inferences about English language learners' content knowledge based on linguistically complex test items.

More research is needed into ways of making academic content more accessible and meaningful to students in bilingual programmes, particularly in areas/subjects considered to be challenging when learning academic content occurs through the second language.

The research findings in respect of 'transfer' tend to support the positive rather than the negative: although more research is needed, the literature points to some evidence for transfer of skills across languages (academic skills, subject knowledge skills, literacy skills).

There is also an urgent need to develop effective bilingual assessment methods that reflect classroom practices of using two (or more) languages for teaching and learning – methods that move away from the notion of monolingual assessment and testing bilinguals as if they were two monolinguals – so that bilingual children are given the opportunity to show their proficiency and competences in both languages.

References

Baker, O. (1996). *Foundations of Bilingual Education and Bilingualism*. 2nd Edition. Clevedon: Multilingual Matters.

Barik, H. & Swain, M. (1976). A longitudinal study of bilingual and cognitive development. *International Journal of Psychology*, **11**, 4, 251–263.

Cambridge IGCSE. (2010). The Cambridge IGCSE Handbook. University of Cambridge International Examinations http://www.cie.org.uk/search?searchfield=handbook%20igcse.

Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, **21**, 617–641.

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Coyle, D. (2006). Developing CLIL: Towards a Theory of Practice. Monograph 6, APAC Barcelona.

Coyle, D. (2007). Content and Language Integrated Learning: Towards a Connected Research Agenda for CLIL Pedagogies. *International Journal of Bilingual Education and Bilingualism*, **10**, 5.

Coyle, D. (1999). Theory and planning for effective classrooms: supporting students in content and language integrated learning contexts. In: J. Masih (Ed.) *Learning through a Foreign Language*. London: CILT.

Cummins, J. (1976). The Influence of Bilingualism on Cognitive Growth. Reproduced in Baker, C. and Hornberger, N. H. (2001) from *Working papers on Bilingualism*, April 1976, 1–43, Ontario Institute for Studies in Education.

Cummins, J. (1980). The Entry and Exit Fallacy in Bilingual Education. Reproduced in Baker, C. and Hornberger, N. H. (2001) from *NABE Journal*, **4**, 25–60.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada. A reassessment. *Applied Linguistics*, **2**, l32–l49.

Cummins, J. (1982). Tests, Achievement and Bilingual Students. Reproduced in Baker, C. and Hornberger, N.H. (2001) from *FOCUS*, February 1982, **9**, 1–7, National Clearinghouse for Bilingual Education, Washington DC, USA.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review*, **56**, 18–36.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society*. Los Angeles: California Association for Bilingual Education.

Cummins, J. & Swain, M. (1983). Analysis-by rhetoric: reading the text or the reader's own projections? A reply to Edelsky et al. *Applied Linguistics*, **4**, 22–41.

Darcy, N. (1946). The effect of bilingualism upon the measurement of the intelligence of children of preschool age. *Journal of Educational Psychology*, **37**, 21–44.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of Language Testing*, Studies in Language Testing 7. Cambridge: UCLES and Cambridge University Press.

Dutcher, N. & Tucker, G.R. (1994). *The Use of First and Second Languages in Education: A Review of Educational Experience*. Washington DC, USA: World Bank.

Edelsky, C. (1990). *With literacy and justice for all: Rethinking the social in language and education*. London: The Falmer Press.

Edelsky, C., Hudelson, S., Altwerger, B., Flores, B., Barkin, F., Jilbert, K. (1983). Semilingualism and language deficit. *Applied Linguistics*, **4**, 1, 1–22.

Garcia, O. (2009). *Bilingual Education in the 21st Century: A global perspective*. Oxford: Wiley-Blackwell.

Grosjean, F. (1982). *Life with Two Languages*. Cambridge, Mass.: Harvard University Press.

Jespersen, O. (1922). *Language, its Nature, Development and Origin*. London: George Allen and Unwin Ltd.

Kelly, V. (1936). Reading abilities of Spanish and English speaking pupils. *Journal of Educational Research*, **29**, 209–211.

Klesmer, H. (1994). Assessment and teacher perceptions of ESL student achievement. *English Quarterly*, **26**, 3, 5–7.

Lambert, W. E. (1977. The effects of bilingualism on the individual: Cognitive and sociocultural consequences. In: P.A. Hornby (Ed.), *Bilingualism: Psychological, social, and educational implications*. 15–27. New York: Academic Press.

McLaughlin, B. (1978). *Second Language Acquisition in Childhood*. Hillsdale, New Jersey: Erlbaum.

Macnamara, J. (1966). *Bilingualism and Primary Education*. Edinburgh: Edinburgh University Press.

Marsh, D. & Lange, G. (2000). *Using Languages to Learn and Learning to use Languages*. Jyvaskala, University of Jyvaskala, Finland: UniCOM.

Martin-Jones, M. & Romaine, S. (1986). Semilingualism: A half-baked theory of communicative competence. *Applied Linguistics*, **7**, 1, 26–38.

Peal, E. & Lambert, W.E. (1962). The Relation of Bilingualism to lntelligence. *Psychological Monographs* **76**, 27, 1–23.

Saer, D. J. (1923). The effects of bilingualism on intelligence. *British Journal of Psychology*, **14**.

Scott, S. (1973). *The Relation of Divergent Thinking to Bilingualism: Cause or Effect?* Montreal: McGill University.

Stewart, J. (2005). Foreign Language Study in Elementary Schools: Benefits and implications for achievement in Reading and Math. *Early Childhood Education Journal*, **33**, 1, 11–16.

Skutnabb-Kangas, T. (1977). Language in the process of cultural assimilation and structural incorporation of linguistic minorities. In: C. C. Elert *et al*. (Eds.), *Dialectology and Sociolinguistics*. 191–199. UMEA: UMEA Studies in the Humanities.

Skutnabb-Kangas, T. & Toukomaa, P. (1976). *Teaching migrant children's mother tongue and learning the language of the host country in the context of the sociocultural situation of the migrant family*. Helsinki: The Finnish National Commission for UNESCO.

Tireman, L. (1955). Bilingual child and his reading vocabulary. *Elementary English*, **32**, 33–35.

Tucker, G.R. (1999). *A Global Perspective on Bilingualism and Bilingual Education*, Carnegie Mellon University, http://www.cal.org/resources/Digest/digestglobal.html

Vygotsky, L. S. (1932). Predislovie [Preface] In: Piaget J. Rech *i m\shieme rebenka* [*The speech and thinking of the child*]. 3–54. Leningrad: Uchpedgt.

Wiley, T. G. (1996). *Literacy and language diversity in the United States*. Washington, DC: Center for Applied Linguistics and Delta Systems.

Appendix A: Comparison of content-obligatory and content-compatible Biology language

Content-obligatory language	*Content-compatible language*
• to describe leaves: 'waxy'; 'spikes'; 'cuticle' • to describe environmental problems: 'deforestation'; 'global warming'; '(bio)degradable'; 'the ozone layer'; 'endangered'; 'fossil fuels'; 'earthquakes'; 'drought' • to describe laboratory experiments: 'test tubes'; 'goggles'; 'pestle and mortar'; 'precipitate'; 'ethanol'; 'iodine'; 'Benedict's solution'; 'control' • to discuss use of fertilisers: 'eutrophication' • to explain the blood system: 'valves'; 'backflow'; '(oxy)haemoglobin'; 'deoxygenated' • to explain plant growth: 'germinate'; 'to wilt' • to describe teeth: 'molars', 'incisors'; 'canines'; 'cusps'; 'dentine'; 'enamel'; 'root' • to identify parts of the human eye: 'cornea'; 'iris'; 'lens'; 'suspensory ligament'; 'yellow spot/fovea'; 'blind spot'	• adjectives or verbs with dependent prepositions: e.g. 'resistant to'; 'suffer from'; 'give off (energy)'; 'react to'; 'respond to'; 'immune to'; 'exposed to'; 'dependent on'; 'protect from'; 'fight off (disease)'; 'adapt to'; 'cut down (trees)'; 'consist of' • phrasal verbs: e.g. 'to break down (a substance)'; 'to carry out (a test)'; 'to set up' (an experiment); 'to speed up' (photosynthesis/a reaction) • verb-adverb collocations: e.g. 'increased exponentially'; 'rises dramatically' • verb-noun and adjective-noun collocations: e.g. 'to have an adverse effect on...'; 'weaken their immunity'

Appendix B: Common Reference levels - Global Scale

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Council of Europe, 2001, p.24

An investigation into the number of special consideration enhancements and their impact on examination grades

Carmen L. Vidal Rodeiro Research Division

Introduction

The GCSE, GCE, Principal Learning and Project Code of Practice (Ofqual, 2011) promotes quality, consistency, accuracy and fairness in assessment and awarding. Therefore, awarding bodies in England need to make sure that candidates have fair access to exams so that they are able to demonstrate their skills and knowledge. Awarding bodies also have to facilitate open access to their qualifications for candidates who are eligible for reasonable adjustments without compromising the assessment of the skills, knowledge or understanding being measured.

A reasonable adjustment is any action that helps to reduce the effect of a disability or difficulty that places the candidate at a disadvantage in the assessment. Reasonable adjustments can be of two types: access arrangements and special consideration. Access arrangements are approved or set in place before the assessment takes place and they constitute an arrangement to give candidates access to the qualification. Examples of access arrangements include: extra time; the use of a scribe; adapting assessment papers, for example providing materials in Braille. Special consideration, the focus of this research, is a post examination adjustment to the marks or grades of a candidate. Applications for special consideration should be submitted by the candidate's school and can be of two types: *present but disadvantaged* or *absent with good reason*.

Present but disadvantaged

Candidates who sat a component/unit are eligible for special consideration if they had been fully prepared and had covered the whole course but performance in the examination or in the production of coursework was affected by adverse circumstances beyond their control. These include:

- temporary illness, accident or injury at the time of the assessment;
- bereavement at the time of the assessment;
- serious disturbance during an examination, particularly where recorded materials are being used;
- accidental events such as being given the wrong examination paper, being given a defective examination paper or tape, failure of practical equipment, failure of materials to arrive on time;
- failure by the centre or awarding body to implement previously agreed access arrangements.

A more exhaustive list of circumstances which might be eligible for special consideration can be found in JCQ (2010).

When candidates were present but disadvantaged, the special consideration enhancements are post examination adjustments to their results. They might cause a relative minor change to the marks obtained in the examination of up to five per cent of the maximum mark for the question paper. The maximum adjustment (or tariff) is reserved for exceptional cases, for example, candidates disadvantaged by the recent death of an immediate family member. However, most adjustments for special consideration are smaller, for example, two per cent of the maximum available mark for candidates with minor illnesses on the day of the examination. It should be noted that a successful application will not necessarily change a candidate's grade.

Absent with good reason

When a candidate has missed a component/unit for acceptable reasons and can produce evidence of that, an adjustment may be made to the overall grade as long as the component/unit was missed in the terminal series and some minimum requirements have been satisfied.

Candidates must have covered the whole course and failure to prepare candidates is not an acceptable reason for an enhanced special consideration grade. In addition, for GCE qualifications, 50% of the total assessment must be completed before a special consideration enhancement may be considered; for GCSE qualifications, 35% of the total assessment must be completed. If too much of the examination has been missed, the candidate will be graded on the marks scored and the certificate will be endorsed to show that not all of the components have been completed.

In the past few years, there have been claims about the number of students receiving extra marks in their examinations due to special consideration increasing year on year (e.g. BBC, 2009; Lipsett, 2009). Also, there has been a great deal of speculation about how pupils and teachers might be abusing the system to boost results, helping schools climb national league tables (e.g. BBC, 2008; Paton, 2009). Therefore, the main aim of this research was to provide evidence in relation to:

- the patterns of special consideration applications
  - over time;
  - by qualification (GCSE vs. A level);
  - by type of school;
- the impact of the special consideration enhancements in the examination outcomes.

Data and methods

Data

The research presents summary statistics of special consideration applications from 2007 to 2009 and detailed analyses of special consideration applications in individual GCSE and A level subjects in the June 2009 session.

At GCSE, eight contrasting subjects were chosen: four subjects that were assessed in a linear fashion (history, geography, mathematics and

religious studies) and four unitised specifications (English, French[1], mathematics and science). At A level, four subjects were chosen: English literature, mathematics, chemistry and history.

GCSE and A level candidates normally take exams from more than one awarding body and therefore might apply for special consideration to one or more awarding bodies. In this research, only candidates who submitted applications for special consideration to the OCR awarding body were considered. GCSE and A level results for those candidates and data on special consideration applications were obtained from OCR's examinations processing system. The data comprised personal details (name, sex, date of birth and school), assessment grade details (session, tier, final mark and final grade) and enhancement details (type of application, outcome and tariff applied).

A measure of students' general attainment (proxy for ability) was computed using data from the National Pupil Database[2]. By assigning scores to the GCSE grades (A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0) it was possible to arrive to a total GCSE score for each student. A 'mean GCSE' indicator was calculated by dividing the total score by the number of subjects attempted. The mean GCSE score was used as a measure of prior attainment for students taking A level subjects. For students taking GCSE subjects, a measure of concurrent attainment was used instead. For each GCSE subject, the concurrent measure was the mean GCSE score calculated excluding the grade in the subject under consideration.

Methods

There are three different types of analyses carried out in this research.

(a) *General statistics on special consideration applications*: Descriptive statistics were used to investigate the patterns in the numbers of special consideration applications over time and by type of qualification.

(b) *Impact of the special consideration enhancements in examination outcomes*:To evaluate the impact of the special consideration enhancements in the examinations outcomes, grades and marks before and after the enhancements were required. Descriptive statistics were then used to calculate the percentages of candidates who certificated in June 2009 and improved their grades due to special consideration.

In order to calculate the number of candidates who improved the overall grade in a subject, applications for special consideration in previous sessions needed to be considered (as GCSE and A level modules could have been taken in different sessions). The analyses were restricted to candidates who certificated in the June 2009 session and had taken any modules used for aggregation in 2008 or 2009 examination sessions. This restriction was made in an attempt to select typical GCSE and A level cohorts.

(c) *Effects of school type on special consideration applications*: To investigate if there were differences at school level in terms of the numbers of special consideration applications, a logistic regression analysis was carried out. Logistic regression is a type of regression analysis that is used when the dependent variable is a dichotomous variable (i.e. it takes only two values, which usually represent the occurrence or non-occurrence of some event) and the independent variables are continuous, categorical, or both. It is used to predict the probability that the 'event of interest' will occur as a function of the independent variables.

In this research, the dependent variable was the request of a special consideration enhancement: the variable took the value 1 if the student applied for special consideration and 0 otherwise. The independent or explanatory variables were the mean GCSE score (proxy for students' ability) and the type of school.

The formal representation of the model was:

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where p was the probability that a student requested special consideration and X_1 and X_2 were the independent variables. β_0, β_1 and β_2 were the regression coefficients, which were estimated from the data.

In this research, the regression coefficients were used to produce estimates of the probabilities of requesting a special consideration enhancement by the candidates' ability and the type of school attended.

Results

The results of the analyses carried out in this research are presented in two sections: section one contains the analysis of special consideration applications where candidates were present but disadvantaged; section two contains the same analyses for candidates who were absent with good reason.

Present but disadvantaged

General statistics

Table 1 presents all the special consideration applications received by OCR (all centres and all qualifications) from 2007 until 2009. These figures show that special consideration applications increased in the period of study (from 78389 in 2007 to 80189 in 2009) and that the majority of the requests were accepted.

Table 1: Numbers and percentages of accepted and rejected special consideration applications (present but disadvantaged), 2007–2009

Session	*Year*	*Accepted*		*Rejected*		*Total number of applications*[3]
		Number	*%*	*Number*	*%*	
January	2007	8757	93.72	202	2.16	9344
	2008	8358	92.88	118	1.31	8999
	2009	9898	88.84	189	1.70	11141
June	2007	62900	91.10	2021	2.93	69045
	2008	71047	93.93	1983	2.62	75639
	2009	64001	92.69	2517	3.65	69048
All	2007	71657	91.41	2223	2.84	78389
	2008	79405	93.82	2101	2.48	84638
	2009	73899	92.16	2706	3.37	80189

Tables 2 and 3 present the numbers of special consideration applications by type of qualification and by tariff in English schools only. Applications for qualifications other than GCSE or A level (e.g. STEP, Entry

1 In this research GCSE French is considered a unitised specification. Although the specification is linear in the sense that all units must be taken in the same session, the entry operates as though it were modular.

2 The National Pupil Database, compiled by the Department for Education, holds national examination data for all candidates who sat an examination in an academic year.

3 'Total' includes applications accepted, rejected, referred to centre or referred to grade review.

Table 2: Special consideration applications (present but disadvantaged) by type of qualification, 2007–2009

Session	Year	GCSE		A level	
		Number of applications	% accepted	Number of applications	% accepted
January	2007	1770	99.10	6639	97.61
	2008	2908	98.93	5294	98.58
	2009	3268	97.95	6378	98.26
June	2007	31361	97.38	30358	96.91
	2008	37298	97.73	32731	96.68
	2009	33628	96.76	29408	95.86
All	2007	33131	97.47	36997	97.03
	2008	40206	97.82	38025	96.94
	2009	36896	96.87	35786	96.29

Table 3: Percentages of approved special consideration applications (present but disadvantaged) by tariff, 2007–2009

Tariff	2007		2008		2009	
	January	June	January	June	January	June
0	2.48	3.89	4.19	3.75	0.30	5.79
1	36.69	25.26	22.96	18.38	15.20	16.04
2	40.88	40.10	42.65	46.03	54.37	39.74
3	9.13	13.14	15.64	13.89	12.55	17.71
4	9.16	13.64	12.18	14.70	14.31	16.45
5	1.66	3.97	2.38	3.25	3.27	4.27

Level, GNVQs) or applications from candidates in schools in Wales, Northern Ireland or Scotland were not included in these analyses.

Looking just at the numbers of applications in Table 2, it seems that similar numbers of requests were submitted for both types of qualifications. However, as a proportion of the entries for each qualification, there were more special consideration requests at A level than at GCSE (for example, 1.35% of GCSE entries requested special consideration in June 2009 vs. 4.52% of A level entries). One reason for this could be the fact that A levels are high stakes examinations and therefore it is more important for candidates to get the 'extra marks'.

Table 3 shows that the most popular tariff applied was 2% of the unit/component total mark, which corresponds to circumstances such as minor illnesses at the time of the examination (e.g. broken limb on the mend, hay fever). Very small percentages of applications were awarded a 5% enhancement.

Individual subjects

The tables presented in this section show summary statistics for special consideration applications in the fourteen GCSE and A level subjects investigated in this research. Detailed analysis for each of the subjects can be found in Vidal Rodeiro (2010).

Results are presented separately for linear and unitised GCSE qualifications. In a modular/unitised qualification a candidate can request special consideration in one or more units and each of these requests counts as one application. In a linear qualification a candidate can request special consideration in one or more papers/components but this counts as one application only.

For individual GCSE and A level subjects, the percentages of special consideration requests, as a proportion of the entries in the subjects,

Table 4: Summary statistics for special consideration applications (present but disadvantaged) in unitised GCSE subjects, June 2009

Subject	Candidates	Candidates (%) with at least one SC application	Candidates (% out of previous column) with overall grade improvement after SC	Candidates (% out of entries in subject) with overall grade improvement after SC
English	46997	1266 (2.69%)	189 (14.93%)	0.40%
French	29696	1268 (4.27%)	106 (8.36%)	0.36%
Mathematics	58697	1853 (3.16%)	115 (6.21%)	0.20%
Science	109953	1766 (1.61%)	81 (4.59%)	0.07%

Table 5: Summary statistics for special consideration applications (present but disadvantaged) in linear GCSE subjects, June 2009

Subject	Candidates	Candidates (%) with at least one SC application	Candidates (% out of previous column) with overall grade improvement after SC	Candidates (% out of entries in subject) with overall grade improvement after SC
History	50621	1932 (3.82%)	314 (16.25%)	0.62%
Geography	35908	832 (1.41%)	126 (15.14%)	0.35%
Mathematics	39467	555 (1.41%)	81 (14.59%)	0.20%
Religious Studies	34262	190 (0.55%)	25 (13.15%)	0.07%

Table 6: Summary statistics for special consideration applications (present but disadvantaged) in A level subjects, June 2009

Subject	Candidates	Candidates (%) with at least one SC application	Candidates (% out of previous column) with overall grade improvement after SC	Candidates (% out of entries in subject) with overall grade improvement after SC
English Literature	7797	709 (9.09%)	25 (3.53%)	0.32%
Mathematics	11499	844 (7.34%)	41 (4.86%)	0.36%
Chemistry	11897	1077 (9.05%)	72 (6.69%)	0.61%
History	12878	1110 (8.62%)	88 (7.93%)	0.68%

were fairly small. Tables 4 and 5 show that, at GCSE, the percentages of candidates with at least one application for special consideration were below 5% for all subjects considered in this research. At A level, the percentages of candidates with at least one application were slightly higher but below 10% (Table 6).

The percentages of candidates with at least one application for special consideration were higher in modular/unitised qualifications than in linear qualifications. Percentages were higher at A level than at GCSE in all subjects considered. It is the case that due to the modular structure of the qualifications, candidates' examinations are spread over a wider period of time (e.g. candidates sit modules on different days, sessions or years), increasing the probability of a temporary illness, injury, or other unforeseen circumstances taking place.

At GCSE, the percentages of candidates improving their grade, as a percentage of the candidates submitting at least one special consideration request, were higher for linear qualifications than for modular qualifications.

In all subjects, both at GCSE and A level, the percentages of candidates out of the total entry who improved their overall grade as a result of a special consideration enhancement were very low (less than 1%).

This research also showed that, in general, candidates in the high attaining groups were more likely to apply for special consideration than those in low attaining groups. At GCSE, in particular, it was more common to improve grades from C to B or from B to A than from D to C, the much debated threshold that determines where a school is ranked in national league tables.

School type analyses

This section investigates the effect of the type of school on the probability of requesting a special consideration enhancement at GCSE and at A level.

Due to the small numbers of applications in each individual subject, all GCSE subjects considered in this research (unitised and linear specifications) and all A level subjects were grouped together. A logistic regression analysis was carried out for each group.

Figure 1 presents the probability of requesting special consideration in GCSE subjects by school type. It shows that candidates in independent schools were more likely to submit a request for special consideration than candidates in state schools[4]. This figure also shows that the probability of applying for special consideration in at least one GCSE unit or GCSE paper/component was very low, ranging from 0.01 to 0.05 (the equivalent to between one and five candidates out of one hundred applying for it).

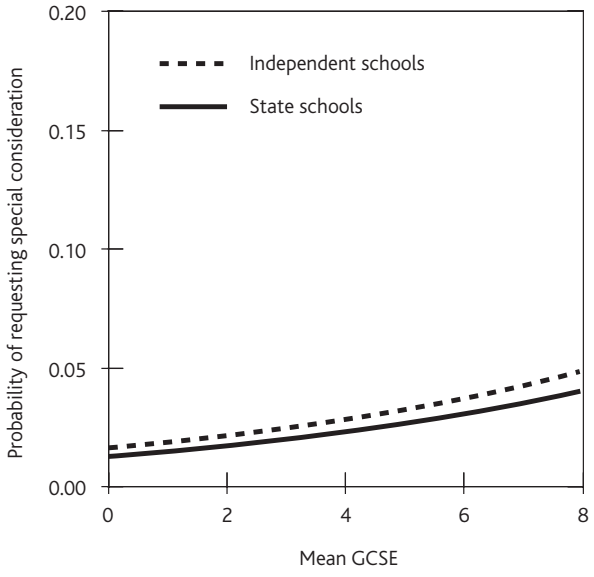


Figure 1: Probability of requesting special consideration (present but disadvantaged) in GCSE subjects by school type

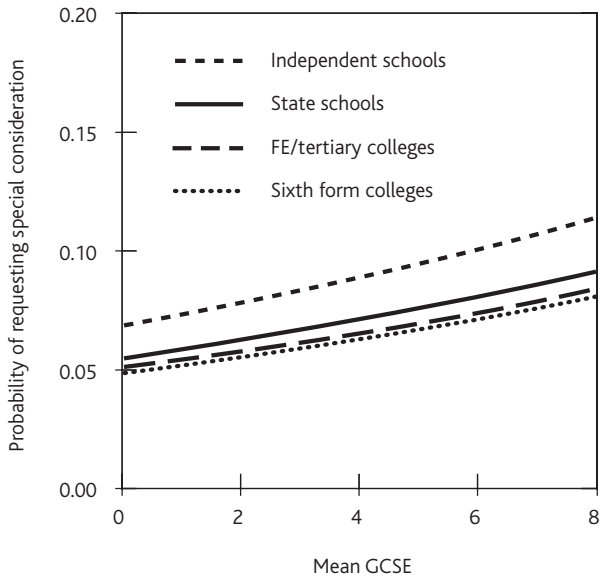


Figure 2: Probability of requesting special consideration (present but disadvantaged) in A level subjects by school type

Figure 2 presents the probability of requesting special consideration at A level by school type and shows that the probability of applying for special consideration in at least one A level unit was higher in independent schools than in any other type of school. The lowest probability was in sixth form colleges.

It should be noted that this probability was also very low (ranging from 0.06 to 0.11) although slightly higher than at GCSE.

Absent with good reason

General statistics

Table 7 presents all the special consideration applications (absent with good reason) received by OCR from 2007 until 2009. It shows that the number of this type of special consideration applications has been increasing in the past few years. Note that the percentages of accepted applications in the January sessions are fairly small. This is probably due to the fact that units/components missed in examination series prior to certification have to be re-entered at a later date.

Table 7: Numbers of special consideration applications (absent with good reason), 2007–2009

Session	*Year*	*Accepted*		*Rejected*		*Total number of applications*[5]
		Number	*%*	*Number*	*%*	
January	2007	42	20.39	6	2.91	206
	2008	60	21.13	2	0.70	284
	2009	61	13.29	394	85.84	459
June	2007	4092	82.68	30	0.61	4949
	2008	4185	83.68	50	1.00	5001
	2009	4857	83.32	856	14.69	5829
All	2007	4134	80.19	36	0.70	5155
	2008	4245	80.32	52	0.98	5285
	2009	4918	78.21	1250	19.88	6288

4 'State' schools include comprehensive schools, grammar schools and secondary modern schools.

5 'Total' includes applications accepted, rejected, referred to centre or referred to grade review.

In 2009 the OCR awarding body received 6288 applications for special consideration where candidates were absent, an increase of about 1000 applications from 2008. Around 80% of the requests were approved. The percentage of approved applications was more than 10% smaller than the percentage of approved applications among candidates who were present but disadvantaged (around 92% in all sessions and years).

Table 8 presents the number of special consideration applications by type of qualification in English schools only. Looking just at the numbers of applications in Table 8, it seems that higher numbers of requests were submitted at GCSE than at A level. However, as a proportion of the unit/specification entries, the percentages of special consideration requests when the candidates were absent were fairly similar for both types of qualifications (e.g. 0.16% at GCSE vs. 0.10% at A level in 2009).

Table 8: Special consideration applications (absent with good reason) by type of qualification, 2007–2009

Session	*Year*	*GCSE*		*A level*	
		Number of applications	*% accepted*	*Number of applications*	*% accepted*
January	2007	43	88.37	3	66.67
	2008	47	97.87	11	90.91
	2009	358	12.85	85	12.94
June	2007	3256	99.45	963	98.75
	2008	3327	99.10	831	97.23
	2009	3997	85.11	1685	81.60
All	2007	3299	99.30	966	98.65
	2008	3374	99.08	842	97.15
	2009	4355	79.17	1770	78.31

Prior to 2009, when candidates missed a unit/component but they were not aggregating in that session, the applications were referred to the centre. This changed in 2009; when OCR issued revised working instructions for special consideration, those applications were instead rejected by the awarding body. This explains the big decreases in the percentages of accepted applications in 2009 shown in Tables 7 and 8.

Individual subjects

Tables 9 and 10 show that, in GCSE subjects, the percentages of candidates with at least one application for special consideration were very small (below 0.50% of the subject entry). In A level subjects, Table 11 shows that the percentages of candidates with at least one application were slightly higher but still below 0.50%.

At GCSE, the percentages of candidates with at least one application for special consideration in modular/unitised qualifications were very similar to those in linear qualifications. Percentages at A level were very similar to those at GCSE.

The percentages of candidates with a missing unit/component who improved their grades after a special consideration enhancement (as a proportion of the candidates with at least one application) were much higher than those of candidates who were present but disadvantaged. The reasoning for this is that when a special consideration enhancement is approved after the candidate missed a unit, an enhanced grade (based on performance on other units/components of the specification) is issued. The adjustment therefore is usually bigger than up to 5% of the total mark in the unit/component missed.

In all subjects, both at GCSE and at A level, the percentages of candidates, out of the total entry, who improved their overall grade as a result of a special consideration enhancement, were fairly low (all below 0.50%).

Table 9: Summary statistics for special consideration applications (absent with good reason) in unitised GCSE subjects, June 2009

Subject	*Candidates*	*Candidates (%) with at least one SC application*	*Candidates (% out of previous column) with overall grade improvement after SC*	*Candidates (% out of entries in subject) with overall grade improvement after SC*
English	46997	164 (0.35%)	127 (77.44%)	0.27%
French	29696	98 (0.33%)	92 (93.88)	0.31%
Mathematics	58697	172 (0.29%)	116 (67.44%)	0.20%
Science	109953	251 (0.23%)	147 (58.57%)	0.13%

Table 10: Summary statistics for special consideration applications (absent with good reason) in linear GCSE subjects, June 2009

Subject	*Candidates*	*Candidates (%) with at least one SC application*	*Candidates (% out of previous column) with overall grade improvement after SC*	*Candidates (% out of entries in subject) with overall grade improvement after SC*
History	50621	134 (0.26%)	103 (76.87%)	0.20%
Geography	35908	124 (0.35%)	83 (66.94%)	0.23%
Mathematics	39467	110 (0.28%)	95 (86.36%)	0.24%
Religious Studies	34262	114 (0.33%)	109 (95.61%)	0.32%

Table 11: Summary statistics for special consideration applications (absent with good reason) in A level subjects, June 2009

Subject	*Candidates*	*Candidates (%) with at least one SC application*	*Candidates (% out of previous column) with overall grade improvement after SC*	*Candidates (% out of entries in subject) with overall grade improvement after SC*
English Literature	7797	32 (0.41%)	29 (90.63%)	0.37%
Mathematics	11499	25 (0.21%)	15 (60.00%)	0.13%
Chemistry	11897	48 (0.40%)	32 (66.67%)	0.26%
History	12878	59 (0.45%)	50 (84.75%)	0.39%

School type analyses

Due to the small numbers of special consideration applications in each individual subject made by candidates who were absent with good reason, a logistic regression analysis was not feasible and an alternative, descriptive, analysis was carried out to investigate the numbers of applications by the type of school.

Tables 12 and 13 present the percentages of schools (as a percentage of the schools registered with the OCR awarding body) with at least one candidate requesting special consideration in GCSE and A level subjects, respectively, in the June sessions from 2007 to 2009.

Table 12: Percentages of schools with at least one GCSE candidate applying for special consideration (absent with good reason), 2007–2009

Year	*Comprehensive*	*Grammar*	*Independent*	*Secondary Modern*
2007	43.94	30.25	16.90	36.02
2008	45.31	28.75	16.62	33.74
2009	47.44	29.30	17.51	36.88

Table 12 shows that around 45% of comprehensive schools offering OCR GCSE examinations submitted at least one application for special consideration; this contrasts with around 17% of independent schools and 30% of grammar schools.

Table 13 shows that there were more sixth form colleges and FE/Tertiary colleges submitting special consideration applications (absent with good reason) than other types of schools. Furthermore, applications in each type of school increased considerably in 2009.

Table 13: Percentages of schools with at least one A level candidate applying for special consideration (absent with good reason), 2007–2009

Year	*Comprehensive*	*FE/Tertiary*	*Grammar*	*Independent*	*Sixth Form College*
2007	14.52	25.63	25.61	15.22	55.64
2008	13.33	28.72	18.29	10.50	47.45
2009	22.28	35.52	30.3	16.25	57.55

Conclusions and discussion

The area of special consideration is complex. A very fine balance is required between allowing candidates, who were disadvantaged for reasons out of their control, enhancements which enable them to be placed on an equal footing with other candidates but not advantaging them to the extent that the assessment objectives of a particular examination are compromised.

It was surprising to find such scarce literature about special consideration, a practice that is fairly common nationally and internationally at secondary school and university. In particular, very little academic writing or research addressing the issue of special consideration in higher education was found (e.g. Croucher, 1995; De Lambert and Williams, 2006; Thompson, Phillips and De Lange, 2006) and there was no academic discussion about this practice in English secondary schools.

Numbers of special consideration applications

The overall picture presented in this report is clear: the numbers of special consideration applications have been increasing in the last few years. Overall applications increased from 83544 in 2007 to 86477 in 2009, while OCR's entries decreased in the same period[6].

There might be a number of reasons for the increases in the numbers of special consideration applications:

- First, as a former chairman of the Office of the Qualifications and Examinations Regulator admitted, "schools are increasingly wise to the rules". In fact, there is more awareness now than in previous years of the special consideration enhancements amongst teachers and parents and more information about the circumstances which might be eligible for special consideration.
- Secondly, it should be noted that the figures reported by the Office of the Qualifications and Examinations Regulator (e.g. Ofqual (2009), Ofqual (2010)) are the numbers of applications for special consideration and not the numbers of candidates receiving an enhancement. The fact that every year the number of modular qualifications increases leads to an increase in the number of applications: in a linear qualification a candidate can request special consideration in one or more papers/components but this counts as only one application; in a modular/unitised qualification a candidate can request special consideration in one or more units and each of these requests counts as one application.
- Thirdly, the increases in applications can be due to increased inclusion, as awarding bodies are committed to meet the needs of those candidates that have been disadvantaged.
- Finally, it could be argued that people are manipulating the system. In fact, there has been speculation about how pupils and teachers might be abusing the system to boost results, helping schools climb in national league tables.

The proportions of approved special consideration requests when candidates were present but disadvantaged, were fairly high (over 90% in most years). However, the percentages of approved applications were about 10% lower for absent candidates. One of the reasons for this might be that units missed in examination series prior to certification had to be re-entered at a later date and applications in those units were rejected even though the candidate might have had a good reason for missing the assessment.

For present but disadvantaged candidates, the research showed that there were more special consideration requests at A level than at GCSE as a proportion of the entries. One reason for this could be the fact that A levels are high stakes examinations (e.g. performance at A level might affect university applications) and therefore it might be more important for candidates to get the 'extra marks'. The research also showed that there were fewer applications for special consideration after missing a time-tabled unit/component for acceptable reasons among A level students than among GCSE students. It could be the case that A level students, due to the high stakes nature of the qualification, were more likely to tolerate unfortunate situations or minor illnesses and do their exams regardless, whereas GCSE students may have been more inclined not to take the exam.

In all ten GCSE subjects investigated in this research, the percentages of present but disadvantaged candidates with at least one application for special consideration were below 5%. At A level, those percentages were slightly higher but below 10% for all subjects. The equivalent percentages for students who missed a time-tabled examination ranged from 0.28% to 0.35% at GCSE and from 0.21% to 0.45% at A level.

There were more applications for special consideration, as a percentage of the entries, in unitised qualifications than in linear ones. This might be partly explained by the fact that with the introduction of modular specifications there are more points in the year when a candidate might have a problem (as examinations are spread over a wider period of time with candidates sitting modules on different days, sessions and even years).

This study also showed marked differences in special consideration applications between schools. Both at GCSE and A level, candidates in

6 Note that to calculate the number of entries, each unit of a modular GCSE or A level subject has been counted individually.

independent schools who were present but disadvantaged were more likely than the same type of candidate in another school to request special consideration. For absent candidates, in GCSE examinations around 45% of comprehensive schools submitted at least one application for special consideration whilst only 17% of independent schools and 30% of grammar schools did so; at A level, there were more sixth form colleges and FE/Tertiary colleges submitting special consideration applications than other types of schools.

Impact of the special consideration enhancements

This research has confirmed that for present but disadvantaged candidates the special consideration enhancements were minor adjustments to their marks, with the most popular tariff applied being 2% of the unit/component total mark (this tariff corresponds to circumstances such as minor illnesses at the time of the examination). Therefore, it was not surprising that the percentages of students who improved their overall grades after a special consideration enhancement were very small: both at GCSE and A level, the percentages of candidates (out of the total entry) who improved their overall grade as a result of a special consideration enhancement were lower than 1%.

It was not unexpected either that the percentages of candidates with a missing unit/component who improved their grades after a special consideration enhancement were much higher than those of candidates who were present but disadvantaged. The reasoning for this is that when a special consideration enhancement is approved after the candidate missed a unit/component, an enhanced grade, based on performance on other units/components of the specification, is issued. The adjustment therefore is usually bigger than up to 5% of the total mark in the unit/component missed.

At GCSE, the percentages of present but disadvantaged candidates improving their grade (as a percentage of the candidates submitting at least one special consideration request) were higher for linear qualifications than for modular qualifications. Percentages for A level candidates were in line with the percentages for modular GCSEs. However, the percentages of candidates who missed a time-tabled unit in a unitised qualification (A levels and new GCSEs) were very similar to those who missed a paper/component in a linear qualification.

Other issues

There has been lots of criticism about how pupils and teachers might be abusing the system to boost results, helping schools climb national league tables, but there is no measure of how frequently such behaviour might occur. However, as shown in this research, the percentages of pupils improving their grades after a special consideration enhancement are so small that this claim seems not to have a strong base.

On the other hand, a survey by Eve and Bromley (1981) revealed that 59% of US college students regarded it as dishonest to feign an illness to avoid taking an examination. It may, therefore, be not too surprising that some students will go to great lengths to avoid or delay taking an examination, or provide evidence to explain a poor performance. In England, claiming special consideration by submitting false information could lead to malpractice.

It might be worth investigating the reverse situation: are deserving students being denied special consideration? There might be a level of abuse which might be justifiable in order to 'rescue' the careers of those worthy candidates whose genuine illness on the wrong day could change the course of their careers.

One of the biggest concerns in relation to special consideration enhancements is the size of the rewards. However, this is a very difficult issue as awarding bodies cannot compromise the assessments and need to be fair with all candidates.

Another concern is related to making judgements on decisions about special consideration applications as there might be a subjective factor when granting an adjustment. The decisions are made by the awarding body based on various factors which are different from one candidate to another. These might include the severity of the circumstances or the date of the examination in relation to the circumstances. Although each case is assessed individually, the best written rules will still require someone to decide on which side of a dividing line each case lies.

References

BBC (2008). *Exam extra marks 'unfair'*. BBC News online, 19 April. Available at http://news.bbc.co.uk/1/hi/uk/7355973.stm.

BBC (2009). *Extra marks for exam day 'stress'*. BBC news online, 24 March. Available at http://news.bbc.co.uk/1/hi/education/7961116.stm.

Croucher, J. (1995). The increasing incidence of special consideration cases at University. *Higher Education Research and Development*, **14**, 13–20.

De Lambert, K. & Williams, T. (2006). In sickness and in need: the how and why of special consideration for students. *Assessment and Evaluation in Higher Education*, **31**, 55–69.

Eve, R.A. & Bromley, D.G. (1981). Scholastic dishonesty among college undergraduates. *Youth and Society*, **13**, 3–22.

JCQ (2010). *Access arrangements, reasonable adjustments and special consideration: general and vocational qualifications*. London: Joint Council for Qualifications.

Lipsett, A. (2009). Significant rise in number of special consideration pupils. *The Guardian*, 24 March. Available at http://www.guardian.co.uk/education/2009/mar/24/special-consideration-exams.

Ofqual (2009). *Statistics for access arrangements and special consideration at GCSE and A level: 2008*. Coventry: Office of the Qualifications and Examinations Regulator.

Ofqual (2010). *Access arrangements for GCSE and GCE: June 2009 examination series*. Coventry: Office of the Qualifications and Examinations Regulator.

Ofqual (2011). *GCSE, GCE, Principal Learning and Project Code of Practice*. Coventry: Office of the Qualifications and Examinations Regulator.

Paton, G. (2009). Students using 'sob stories; for extra marks. *The Telegraph*, 24 March. Available at http://www.telegraph.co.uk/education/secondaryeducation/5043442/Students-using-sob-stories-for-extra-marks.html.

Thompson, P., Phillips, J. & De Lange, P. (2006). The assessment of applications for special consideration: a conceptual framework. *Accounting education: an international journal*, **15**, 235-238.

Vidal Rodeiro, C.L. (2010). *Special consideration: a statistical investigation on the number of requests and the impact of the enhancements*. Research Report. Cambridge: Cambridge Assessment.

The effect of manipulating features of examinees' scripts on their perceived quality

Tom Bramley Research Division

Introduction

Expert judgment of the quality of examinees' work can play an important part in several assessment contexts. First and most obvious is the *marking* (scoring) of a response to a constructed-response item or an open-ended item. Here the task of the judge[1] is usually to assign a number (the 'mark') to the response according to guidelines or instructions in the mark scheme (scoring rubric). A second aspect is *standard-setting* – deciding on a cut-score on the score scale that represents the boundary between two categories such as pass/fail, grade A/grade B, advanced/proficient. Here the task of the judge(s) is to decide whether the quality of examinees' work at a particular mark point on the scale is worthy of the higher or lower categorisation, usually with reference to explicitly defined performance standards. An example of this is the 'Body of Work' standard-setting method described by Kingston, Kahl, Sweeney and Bay (2001). A third, and closely related, aspect is *standard-maintaining* – deciding on a cut-score that represents the same performance standard as equivalent cut-scores that have been set on previous versions of the test. Here the task of the judge(s) is to find the point on the score scale where the quality of examinees' work matches that of examinees at the same boundary on previous versions of the test. The mandatory procedures for setting grade boundaries on high-stakes school examinations in England and Wales (GCSEs and A levels[2]) include this kind of judgment as one source of evidence amongst others to be considered in an 'award meeting' (Ofqual, 2009, p.37). A fourth aspect is *comparability monitoring* – comparing the quality of examinees' work on different tests where for whatever reason it is deemed important that performance standards are comparable. This is a very broad area that will have a different focus in different international contexts. An example from England would be comparing the standard of work produced by examinees at the grade A boundary in a particular GCSE or A level subject from assessments produced by different examination boards (awarding bodies). Reviews of comparability methods involving expert judgment used in the UK can be found in Adams (2007) and Bramley (2007).

Standard setting, standard maintaining, and comparability monitoring all have in common that the judge's task is to make a holistic judgment about the quality of examinees' work (henceforth referred to as a 'script'), either at the level of an examination paper, or at the level of a complete assessment (which might involve several papers as components).

1 The expert making the judgment is generically referred to as a 'judge' in this article. Other more context-specific terms include marker, rater, examiner and awarder.

2 General Certificate of Secondary Education (GCSE) examinations are taken in England and Wales at age 16+ at the end of compulsory schooling, Advanced Subsidiary (AS) levels are taken at 17+, and Advanced (A) levels at age 18+ (this second year of post-compulsory examinations being referred to as A2).

It has frequently been found, in a variety of contexts and using a variety of methods, that *holistic* judgments of the relative quality of scripts (made in the absence of knowledge of the mark totals) do not correspond exactly to the ordering of the scripts by their mark totals (e.g. Bramley, Bell and Pollitt, 1998; Gill and Bramley, 2008; Baird and Dhillon, 2005; Edwards and Adams, 2002; Jones, Meadows and Al-Bayatti, 2004). Indeed, the finding is often made even in contexts where the judges *are* aware of the mark totals, such as traditional grade awarding meetings for GCSEs and A levels. Here it is not unusual for a script with a lower total mark to be judged more worthy of a higher grade than a script with a higher total mark – although the nature of the award meeting ensures that these 'reversals' are far less common and of lesser magnitude than in exercises where the judges do not know the mark totals.

It is therefore of great importance to understand in as much depth as possible the factors that influence these holistic judgments in order to have confidence in the outcomes of exercises that use them. It seems likely that there would be several, perhaps many, different features of the scripts that influence the judgments, and that some of these might be deemed to be more or less valid than others. For example, if 'quality of handwriting' was found to be a factor, this would (presumably) not be considered a valid cause of perceived difference in quality. It also seems likely that there would be differences among judges as to which features were more relevant to their own decisions.

Several different methods have been used to get at the underlying causes of the judges' decisions. The most obvious method is simply to ask the judges what factors they thought were most relevant to their judgments. This has been done in many inter-board comparability studies (e.g. Edwards and Adams, 2002; Fearnley, 2000; Jones *et al*., 2004). The advantage of this method is its transparency, but there are several disadvantages. First, it is not possible to know whether the judges are correct – that is, whether they are actually aware of the factors underlying their judgments. This is the general problem of reliability of self-report measures, discussed in several sources (e.g. Nisbett and Wilson, 1977; Leighton, 2004). Second, it seems likely that judges would avoid mentioning any obviously invalid factor, such as handwriting, in case it cast doubts on their expertise. Third, it is often the case that the judges report something that is rather hard to pin down precisely, such as 'depth of understanding'. Finally, it is not possible to determine the relative importance of the factors that judges report.

A second method is to try to discover the cognitive processes underlying the judges' judgments, and the features of the scripts that they are attending to, by verbal protocol analysis (Ericsson and Simon, 1993). Here, judges are asked to 'think aloud' as they make their judgments and the transcripts of their verbalisations are coded and analysed. Examples of this approach can be found in Crisp (2008a, b),

Suto and Greatorex (2008), and Greatorex and Nádas (2008). The advantages of this approach over the previous one are that it gets closer to the actual decision-making, and avoids post hoc rationalisation (or invention). Some of the same disadvantages apply – for example, it is still not necessarily the case that features elicited this way are in fact the most causally relevant to judges' decisions.

A third approach would be to carry out post hoc analysis of scripts that have been involved in a judgmental exercise, comparing scripts with the same total score that were judged to be of different quality and attempting to identify points of difference between them that might have been responsible for the perceived difference. One disadvantage of this approach is that scripts from different examinees can differ on many different features and it would be difficult to determine which features had been relevant to the judgments.

A fourth, related, approach would be to identify, a priori, features of scripts that might be salient to judges. Each of a set of scripts could then be rated on the presence or absence of these features (or the degree to which they possess them). Then the relationships between the coded features and perceived quality could be analysed. Potential problems with this approach include multi-collinearity (similar types of feature tending to cluster together), separating causation from correlation, and the risk of discovering spurious associations. But further cross-validation work could minimise these problems. An example of this approach can be found in Suto and Novaković (*in press*).

A fifth approach, and the one tried in this study, is to carry out a controlled experiment, preparing different versions of the same scripts that differ only in a single feature while keeping others constant, in particular the total score. Differences between the versions in perceived quality can then be attributed to the changes made. The advantage of this approach is that it offers a rigorous way to isolate the effect of different script features on perceived quality, and thus allow stronger causal conclusions to be drawn. A disadvantage is that the features have to be specified in advance – so potentially could be found to be not relevant (although this does avoid the pitfall of capitalising on chance associations in a post hoc analysis). A further disadvantage is that only a small number of features can be tested in one experiment – thus leaving the possibility that other, untested, features would be found to be of greater importance. A final disadvantage is that the method lends itself best to features that can be easily manipulated experimentally. These disadvantages notwithstanding, the approach seems promising and, to the author's knowledge, has not been tried before. This study therefore represents a new approach to this difficult problem.

The particular judgmental method used in this study was the rank-ordering method for standard maintaining (Bramley, 2005; Black and Bramley, 2008; Bramley and Black, 2008). A detailed description of this method is beyond the scope of this article, but it is essentially an extension of Thurstone's (1927) method of paired comparisons. Each judge's task is to put sets of scripts (with mark totals removed) into rank order according to perceived quality. The key features of the method are: i) that it involves *relative* rather than absolute judgments, so scripts are compared with each other rather than with a nominal standard. This allows any differences among the judges in personal (absolute) standards to cancel out; and ii) the analysis of the rankings with a latent trait (Rasch) model locates each script on a scale of 'perceived quality' which can then be related to the total score scale. The rank-ordering method has been used in a variety of settings and is evaluated in Bramley and Gill (2010).

Method

The examination paper chosen for the study was one unit from a GCSE Chemistry examination, from June 2007[3]. This examination had a good mix of questions requiring different types of response, and it had been marked on-screen, so both the scanned images of the scripts and item level data (the marks of each examinee on each sub-question) were available. There were 39 sub-questions on the paper and the maximum possible score was 60. Examinees wrote their answers to the questions in allocated spaces on the question paper.

Features to be manipulated

Four features of scripts were chosen to be manipulated in this study. They were chosen because they were hypothesised to be relevant to perceived quality, because they could be relatively easily manipulated, and because they were not too subject-specific (meaning that it might be appropriate to generalise results to other situations).

1. Quality of written English. Some of the questions on the paper required two or more lines of writing in the response. The quality of the writing in terms of surface features such as spelling and punctuation could conceivably have an effect on the perceived quality of a script - those with better writing being perceived to be better. It was expected that the judges, as professional Chemistry examiners, would probably not be influenced by this feature and that it could therefore serve to aid interpretation of the sizes of any other effects that were found.

2. Missing response v incorrect answer. When judges compare two scripts with the same total score, are they more likely to be impressed by an examinee who has attempted all the questions, even if they have a lot of incorrect answers, or is a script containing fewer incorrect answers but a higher number of missing responses perceived more favourably? No hypothesis was made about the direction of this effect.

3. Profile of marks in terms of good fit to the Rasch model. If an examinee's set of responses fits the Rasch model, then they should have gained more of their marks on the easier questions and fewer marks on the harder questions. On the other hand, a misfitting examinee with the same total score will have picked up more marks than expected on the harder questions, but these will be counterbalanced by some lower marks than expected on the easier questions. It was hypothesised that judges might be more impressed by the performance of a misfitting examinee than a well-fitting examinee with the same total score. Anecdotal impressions and observations have suggested that examiners are more likely to take a good answer to a hard question as evidence of high ability (rather than, for example, cheating, special knowledge or good luck), and more inclined to treat a poor answer to an easy question from such an examinee as evidence of carelessness rather than low ability. This impression can be further supported by an analogy with high jumping – someone who clears a high bar but knocks off a low bar might (arguably) seem to be a better jumper than one who clears the low one but not the high one. This feature has similarities with the 'consistency of performance' investigated by Scharaschkin and Baird (2000), but whereas they defined consistency in terms of the range of observed question marks, the fit measure used

3 OCR (Oxford, Cambridge and RSA Examinations) is a UK awarding body. The Chemistry examination paper used in this study was OCR's GCSE Chemistry (Gateway) Higher Tier, unit code B641. It can be downloaded from http://www.ocr.org.uk/Data/publications/past_papers_2007_june/GCSE_Gateway_Chemistry_B_B641_02_June_2007_Question_Paper.pdf . Accessed 21/4/09.

here takes account of variability in question difficulty.

4. Profile of marks in terms of answers to 'good chemistry' questions. Although all sub-questions on a chemistry paper could be said to be testing chemistry by definition, it seems plausible that some sub-questions might conform more to a purist's idea of what chemistry is than others do. It was hypothesised that judges would be more impressed by an examinee who had gained a higher proportion of their marks on the 'good chemistry' questions than an examinee (with the same total score) who had gained a higher proportion of marks on the other questions. Although at one level this feature is obviously specific to the paper, it seems plausible that the concept could generalise – that is, it may be that on maths papers expert judges are particularly influenced by performance on questions that bring out the 'good mathematicians', or on language papers the 'good linguists' etc.

Script selection

1000 scripts were initially selected. 250 were sampled uniformly across the mark range from total test scores of 11 to 50 (out of 60), five scripts on each mark point. The other 750 were sampled at random. This was to ensure that there would be enough scripts to select from at each mark point. The data from these 1000 examinees were then analysed both with classical item analysis and the Rasch model[4] in order to obtain indices of item difficulty, omit rate, and person fit. Ten scripts were then chosen for each feature to be manipulated, giving a total of 40 scripts. The manipulations made in each category are described below.

1. Quality of English. Ten scripts were chosen from across the mark range. 13 sub-questions on the question paper were identified where the space for the examinee's answer had two or more lines. The responses of each examinee to these sub-questions were changed (where possible) to improve the spelling, grammar and punctuation. It is important to stress that these changes were relatively slight and superficial. No change was made that might have changed the mark awarded to the response. In the few cases where the examinee's response was too incoherent to 'improve' without risking altering the mark it would have obtained, it was left alone.

2. Missing response v incorrect answer. Ten scripts were chosen from across the mark range. Five scripts were chosen because they had a high number of blanks (missing responses) and five because they had a low number of blanks. For the five scripts with a high number of blanks, incorrect responses to the sub-questions that had been left blank were located from other examinees with the same total mark (using the 950+ non-selected scripts). It was thought important to use examinees with the same total mark to supply the incorrect responses because their responses would be more likely to be typical of what the original examinee might have written. For the five scripts with a low number of blanks, sub-questions that might plausibly have been left blank were identified (based on difficulty, position in paper, and the overall omit rate). At the high end of the mark range the manipulation changed the response to about four sub-questions out of a total of 39 sub-questions on the paper. At the low end of the mark range as many as 13 responses were changed.

3. Profile of marks in terms of good fit to the Rasch model. Ten scripts were chosen from across the mark range. Five scripts were chosen because the examinee had a high value for the misfit statistic (indicating a 'misfitting' examinee); and five because the examinee had a low (high negative) value for the fit statistic (indicating an 'overfitting' examinee). For the five misfitting examinees, the fit statistics for each sub-question were inspected to discover where the misfit lay – that is, which of the easier sub-questions they had got unexpectedly low marks on, or which of the harder sub-questions they had got unexpectedly high marks on. Responses from examinees (with the same overall total score) who had obtained a more expected score on these sub-questions were located in the (950+) remaining unused scripts. Care was taken to ensure that the number of marks to be gained on the easier sub-questions was balanced by the number of marks to be lost on the harder sub-questions so that the manipulation did not change the overall total score. For the five 'overfitting' examinees, the opposite was done – that is, responses were located from the remaining unused scripts that would make their profile fit less well, again taking care to ensure that marks gained equalled marks lost. This was done in a plausible way – that is, not for example by making the easiest question wrong and the hardest right, but by altering responses to sub-questions in a range of difficulties closer to the examinee's ability estimate. In all cases the manipulation involved changing each examinee's response to about ten sub-questions on the paper.

4. Profile of marks in terms of answers to 'good chemistry' questions. An examiner who had set papers for the same suite of examination papers, but who was not an awarder for this particular paper, was recruited to identify the 'good chemistry' sub-questions. He identified 20 sub-questions worth 30 marks in total. Each examinee's total on the 'good chemistry' and 'non-good-chemistry' sub-questions was calculated. Ten scripts were chosen from across the mark range. Five scripts were chosen because the examinees had scored a high proportion of their marks on the 'good chemistry' sub-questions; and five because the examinees had scored a low proportion of their marks on these sub-questions. For each set of five, responses from the remaining pool of unused scripts were used to change the balance of marks in the appropriate direction. As before, care was taken to find replacement responses from examinees with the same (or if this was not possible a very similar) overall total score, but a further precaution was taken – namely not inadvertently to change the mark profile in terms of Rasch fit. This was achieved by making sure that the marks gained and the marks lost were from 'good chemistry' and 'non-good-chemistry' sub-questions that were matched in terms of difficulty. In all cases the manipulation involved changing each examinee's response to about ten sub-questions on the paper.

Script preparation

It was important to ensure that the original and manipulated versions of each script were written in the same handwriting (in order to rule out handwriting as a potential feature influencing the comparison). It was also important to ensure that the 40 pairs of scripts (original + manipulated) were written in *different* handwriting (so they looked like 40 different examinees), and to ensure that all handwriting looked as though it could plausibly have been produced by 16 year olds.

To this end, the author's colleagues volunteered (or were persuaded) to act as 'scribes', and produce a pair of scripts for the study. They did this by copying out the original answers onto a blank question paper, and then produced the manipulated version by copying out onto a second

4 Missing responses were scored zero, as they were in the actual examination.

5 The 'residual fit' statistic in Rumm2020 (Rumm Laboratory, 2004).

blank question paper the original answers plus the necessary changes. The scribes were told not to try to imitate the examinees' handwriting, but to use their own style, adapted to make it more like a 16 year old's (only if absolutely necessary). The scribes were also asked to reproduce all the crossings out, mis-spellings, diagrams etc. in order to make the scripts look as authentic as possible. The only feature they did not copy was the number of words per line (which is very dependent on size and spacing of handwriting).

The scripts were then given a front page containing a random two-letter ID to be used in the study. Each script was then scanned, thus creating a set of 80 pdf documents that could be printed out as many times as required by the design of the study.

Judges

The expert judges invited to take part in the study were the six members of the awarding panel (the group of experts responsible for standard maintaining) for this Chemistry paper in June 2007. All agreed to take part. Before attending the meeting, the judges were asked to carry out some preparatory work. The purpose of this was to ensure that they were fully re-familiarised with everything relating to this particular examination. They were sent a package of advance materials containing: i) the question paper; ii) the mark scheme; iii) the specification grid; iv) the item level data analysis report; v) the report on the examination prepared by the Principal Examiner; and vi) two examinees' scripts (not from the study) to re-mark.

They were asked to read all the material before attending the meeting, and to re-mark the two scripts (so they could re-orient themselves to the kinds of responses examinees had given). The aim was to ensure that the judges would be as well-prepared as possible to make the rank-ordering judgments required of them.

Design

The design of the study was necessarily complex. The aim was to ensure that each judge made a judgment about each script in the study. However, the intention was to conceal from the judges the fact that there were two versions of each script, in case that knowledge influenced the outcome.

Each pack of scripts to be ranked contained four scripts. The packs for each judge were arranged into two sets of ten. Across the first ten packs, each judge saw one version of all 40 scripts in the study. The first pack contained scripts from the top end of the mark range, going down to the tenth pack which contained scripts from the bottom end of the mark range. Across the second ten packs, each judge saw the other version of each script, i.e. the version (original or manipulated) that they had not seen in the first ten packs. The second ten packs also ran from the top end (pack 11) to the bottom end (pack 20) of the mark range.

Each judge saw a different selection of scripts in each pack, and whether they first saw the original or the manipulated version of each script was randomised. The average mark range of the scripts in each pack was around five marks, that is, the best script in each pack had usually received a test total score five higher than the worst script in each pack, although the random nature of the allocation algorithm meant that some packs had wider and some had narrower ranges than this.

Instructions to judges

At the start of the meeting, the judges were given some general background to the study. This information was presented orally. The purpose of the exercise was presented as being to discover what features of scripts influence judgments of relative quality when scripts are put into rank order. The main contrasts of this study with a conventional award meeting were highlighted: i) relative rather than absolute judgments; ii) judgments of scripts across the whole mark range rather than at a particular grade boundary; and iii) no marks visible on the scripts.

The specific instructions were then given to the judges on paper (see Appendix A), and these were then explained. All relevant information about the purpose and the mechanics of the study was given to the judges with the one exception mentioned above – they were not told that there were two versions of each script. They were told that the second set of ten packs contained the same scripts that they had seen in the first ten packs, but in different arrangements (i.e. shuffled differently among the ten packs). While it was true that the arrangement was shuffled, it was also the case that each script they saw in the second ten packs was a different version of the script they had seen in the first ten packs. In order to facilitate this subterfuge, the scripts had been given random 2-letter IDs (e.g. 'DL') in the hope that these would be so unmemorable that the judges would not be aware that the IDs of the scripts in their second ten packs were different from those in the first ten packs (which had been cleared away before judgments on the second ten packs began).

The judges were asked to work independently, and to refrain from making tied rankings. They were allowed to indicate any scripts they felt were genuinely of the same quality by placing a bracket around them on their record sheets. Past studies have found that this helps judges to move on, and avoid getting 'hung up' on difficult judgments. It was emphasised to the judges that their rankings should be based on overall holistic judgments of quality, using all the kinds of information that they would normally consider in an awarding situation, and that they must not re-mark the scripts.

The final part of the meeting involved collecting written answers from each of the judges to questions that were designed to elicit their opinions on the features of the scripts that they thought influenced their judgments, and what they expected the outcomes of this study to be. After collecting the written feedback, the full purpose of the study (including the existence of two versions of each script) was revealed to the judges in a final plenary discussion session.

Results

Scale evaluation

The recording sheets contained 20 sets of rankings of four scripts for each of the six judges. These data were double-keyed into a spreadsheet and checked. The data were analysed using a Rasch formulation of Thurstone's paired comparison model (see Andrich, 1978a; Bramley, 2007). The paired comparison model requires the rankings to be converted to sets of paired comparisons. Each ranking of four scripts yields six paired comparisons. The model fitted was:

$$\ln\left[\frac{p(i>j)}{p(j>i)}\right] = B_i + B_j$$

where $p(i>j)$ is the probability that script i is ranked above script j, and B_i and B_j are the 'measures' of perceived quality for scripts i and j respectively.

FACETS software (Linacre, 2005) was used to fit this model. The full

FACETS output is given in Appendix B. No script was ranked first or last in every pack in which it appeared, so measures could be estimated for all 80 scripts. The separation reliability index (analogous to Cronbach's Alpha) was high at 0.98, showing that the variability in perceived quality among the scripts could not be attributed to chance. The fit statistics for both scripts and judges showed a slight tendency towards over-fit suggesting that the judges were perceiving the trait in the same way and that there was less variability in their judgments than modelled. All these scale statistics need to be treated with some caution because the paired comparison analysis, when derived from rankings, violates the assumption of local independence between paired judgments. However, there was no indication of any serious problems with the scale[6].

It was of great interest to see how the measures of perceived quality related to the marks awarded to the scripts, which the judges were completely unaware of when making their judgments. A low correlation would suggest that the judges were perceiving a different construct of quality than that resulting from the application of the mark scheme.

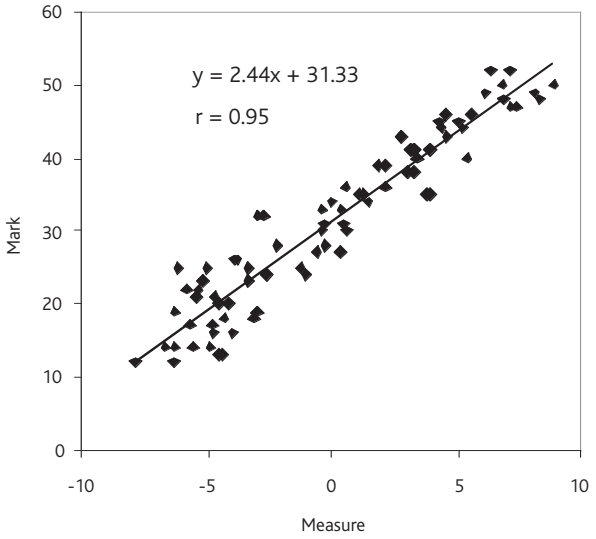


Figure 1: Plot of mark (total score) of script against measure of perceived quality

Figure 1 shows that there was a very high correlation (0.95) between the marks and the measures. This is evidence of the expertise of the judges and the validity of the mark scheme. The slope of the linear regression of mark on measure, 2.44, gives an approximate 'rate of exchange' mapping the scale of perceived quality (in logits[7]) into the mark scale. Since the choice of regression line is itself somewhat arbitrary (Bramley and Gill, 2010), and a standardised major axis has a slope of 2.56, it seems reasonable to take a rough conversion factor of 1 logit = 2.5 marks for interpreting effect sizes.

Effect of experimental manipulation on perceived quality

For the analyses reported below, the 20 scripts in each category were grouped into ten pairs according to the research hypotheses about the effect of the experimental manipulation on perceived quality. Figures 2 to 5 show the differences between the measures obtained by the scripts in the original and manipulated versions. Scripts perceived to be of exactly

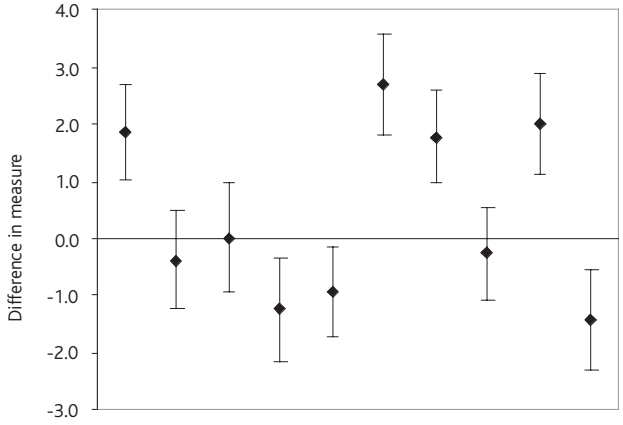


Figure 2: Plot of difference between measures of scripts with improved quality of English and measures of original scripts

the same quality in both versions would have a value of zero for this difference. The error bars show ±1 standard error of measurement (calculated as $(se_1^2 + se_2^2)^{1/2}$.)

Figure 2 appears to show no consistent effect of changing the quality of English – some scripts had a higher measure in the improved version (points above the x-axis line) and some in the original version (points below the x-axis). The biggest differences were all in the 'improved' direction, however, which is not too surprising.

Figure 3 shows that scripts with incorrect answers were fairly

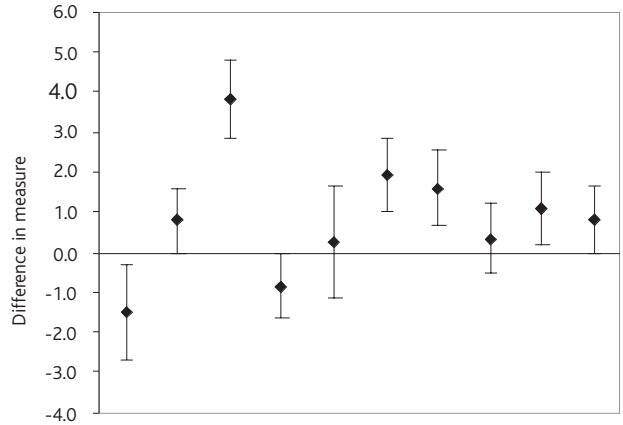


Figure 3: Plot of difference between measures of scripts with incorrect answers and measures of scripts with missing answers

consistently perceived to be of better quality than those with missing answers. Eight of ten points were above the x-axis. No directional hypothesis had been made about whether the missing or incorrect answers would be perceived to be better.

Figure 4 appears to show no consistent effect of changing the degree of fit, but the biggest differences were clearly in favour of worse fit, as hypothesised.

Figure 5 shows that scripts with a higher proportion of good chemistry marks were fairly consistently perceived to be of better quality than those with a lower proportion, as hypothesised.

The above graphs have illustrated the main findings, and shown that the effect of changing the scripts was in the direction predicted by the research hypothesis (where there was a directional hypothesis). However, it appears from the graphs that none of the effects was particularly large

6 A parallel analysis of the rankings was carried out using the Rasch Rating Scale Model (Andrich, 1978b). The resulting measures of perceived quality correlated 0.999 with those from the paired comparison model. The separation reliability index was the same (0.98).

7 The logit (log-odds unit) is the arbitrary unit created by the analysis method.

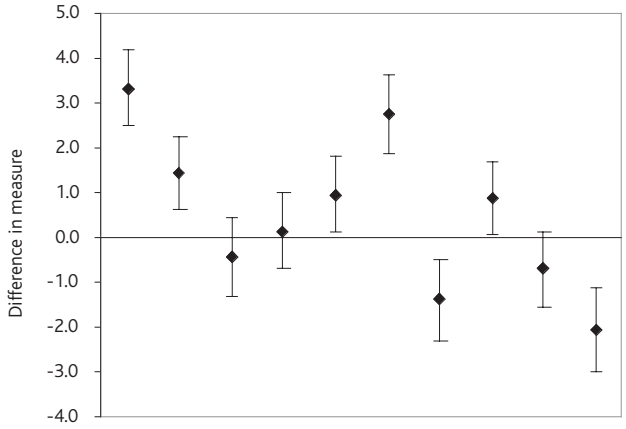


Figure 4: Plot of difference between measures of scripts with worse fit to the Rasch model and measures of scripts with better fit

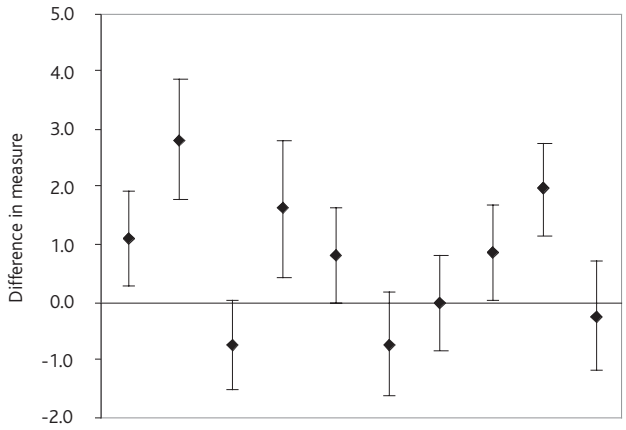


Figure 5: Plot of difference between measures of scripts with a higher proportion of 'good chemistry' marks and measures of scripts with a lower proportion of 'good chemistry' marks

– although given that each pair of scripts would have received the same total score it was not expected that large differences would be found.

One way of analysing the effect of the manipulations is to carry out a 'fixed effects' chi-square test of homogeneity (Cochran, 1954; Linacre, 1992). The 'common effect' of the manipulations in each category is calculated as the information-weighted average of the ten differences, with associated standard error. A z-test then shows whether the common effect is measurably different from zero. A chi-square test can then test the hypothesis that all ten differences are statistically equivalent to one common 'fixed effect', apart from measurement error. The results of this analysis are shown in Table 1. For each category, the results are presented in terms of positive differences. So the 'English' category shows the effect of improving the quality of English; the 'Zeros' category shows the effect of having incorrect answers instead of missing responses; the 'Fit' category shows the effect of having a more misfitting profile of marks in terms of the Rasch model; and the 'Chemistry' category shows the effect of having a higher proportion of marks on the 'good chemistry' questions. Table 1 includes a further column that tries to quantify the effect of the manipulations in a more meaningful way – that of the original raw score scale. This was done by converting logits to marks using the conversion factor of 2.5 described previously.

Table 1 shows that all manipulations, except that of improving the quality of English, had an average (common) effect that was measurably

Table 1: Tests of significance of manipulations in terms of 'common effect' and test of homogeneity

Category	*N*	*Common effect (logits)*	*Standard error (logits)*	*Common effect (marks)*	*Common effect*		*Homogeneity*	
					z	*p-value*	χ^2	*p-value*
English	10	0.42	0.27	1.1	1.55	0.061	28.92	0.001
Zeros*	10	0.85	0.29	2.1	2.91	0.004	20.28	0.016
Fit	10	0.54	0.27	1.4	2.00	0.023	34.61	<0.001
Chemistry	10	0.64	0.28	1.6	2.28	0.001	14.94	0.092

*The 'Zeros' category uses a 2-tailed test of whether the common effect is significantly different from zero; the other categories use 1-tailed tests because of the directional hypothesis.

greater than zero, using a criterion of a p-value for the common effect being less than 0.05. (The value for quality of English was very close to meeting this criterion). However, the homogeneity tests showed that the hypothesis that the effect of the manipulation was constant across all ten scripts could be rejected for all manipulations except that of the proportion of marks gained on 'good Chemistry' questions – and even this was close to being rejected. The largest effect (2.1 marks) was the difference between scripts containing wrong answers as opposed to missing responses. The effects of 'good Chemistry' and 'more misfit' were around 1.5 marks, but only slightly higher than the effect of improving the quality of English (1.1 marks).

Judge feedback

For reasons of space, it is not possible to describe the judges' responses in detail here, but in summary, their comments provided a lot of support to the experimental findings. All six judges seemed to endorse the idea that answers to the 'good chemistry' questions would be influential in their judgments, and this was indeed found. Five of the judges also endorsed the idea that good answers to difficult questions outweigh poor responses on easy questions. As hypothesised, the wrong answers on easy questions can be attributed to 'slip-ups' when making holistic judgments of quality. Interestingly, there were differences among the judges in their thoughts on how missing responses would affect their perception. Two of the judges said that blanks give a worse impression than wrong answers, but another two judges suggested the opposite. This was as hypothesised – there was no directional hypothesis for this effect because both seemed plausible. However, the analysis of rank-order judgments clearly suggested that the scripts in this study that had blanks instead of incorrect answers were perceived to be of lower quality, and this was the largest effect found. There was some agreement among the judges that the judgments ought not to be influenced by the quality of English, yet also some recognition that in practice it might be hard to ignore. The point was made that poor English can also hinder the communication of the examinee's knowledge.

Discussion

The main finding of this study was that it is not only the total score, but also where and how the marks have been gained that contributes to perceived quality. The most influential feature of scripts in determining their perceived relative quality was the presence of blank (missing) answers. Scripts with these were perceived to be worse by the equivalent

of 2 marks than scripts with the same total score that had incorrect but non-missing answers. It should be emphasised that all the experimental manipulations made in this study *did not affect the total score of the script*. Increasing the proportion of marks gained on questions testing 'good Chemistry' or the proportion of marks gained on more difficult questions also increased the perception of quality.

The recognition that the profile of marks contributes to perceived quality is implicitly recognised when setting grade boundaries in GCSEs and A levels where the expert judges are sometimes directed to focus on performance on questions known as 'key discriminators'. Different questions might be deemed to be 'key discriminators' for different grade boundaries. Although 'good chemistry' is not the same concept as a 'key discriminator', there is the same idea that a holistic judgment can be based on a particular *subset* of the total performance, or that different parts of an examinee's performance can carry more weight.

The implication is that the decision of grade-worthiness (in an award meeting), or of relative quality (in a rank-ordering exercise) is dependent to a large extent on the internal profile of marks in the scripts chosen to represent all the scripts at a particular mark point. There is thus something of a tension between the rationale of judgmental standard maintaining exercises (in awarding meetings or rank-ordering exercises) and the purpose of grading. Applying a grade boundary to a mark scale ensures that everyone with a total mark on or above the boundary (up until one mark below the next boundary) receives the grade. How an examinee has achieved their total score is irrelevant – any mark profile that yields the same total will receive the same grade. However, the mark profile has been shown to be important in the judgmental standard-maintaining. If this were the only thing determining the grade boundary, it (the boundary) would be affected by the particular scripts chosen for scrutiny by the judges.

While the mark range of scripts considered at an award meeting is closely controlled, there is (as yet) no such control exercised over the profile of marks within the scripts or of other features, such as the number of missing responses, when selecting scripts for scrutiny. This study suggests that it might be possible to improve the validity of an awarding meeting (or rank-ordering exercise) by choosing scripts that are representative of all scripts on that mark point in terms of the features that this study has shown do influence judgments. With the increasing availability of item level data, this is now a possibility in a wide range of examinations.

Future work could attempt to replicate the findings here, ideally with more scripts in each category, and to explore the extent to which they can be generalised to subjects other than GCSE Chemistry. It seems reasonable to hope that similar results would be obtained in other examinations with similar types of question. It will also be interesting to identify and test other potential features for experimental manipulation.

References

Adams, R. (2007). Cross-moderation methods. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. 212-245. London: Qualifications and Curriculum Authority.

Andrich, D. (1978a). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, **2**, 449-460.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika*, **43**, 4, 561-573.

Baird, J.-A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact*. AQA Research Report RPA_05_JB_RP_077. Guildford: AQA.

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357-373.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, **6**, 2, 202-223.

Bramley, T. (2007). Paired comparison methods. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. 246-294. London: Qualifications and Curriculum Authority.

Bramley, T., Bell, J.F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research & Perspectives*, **25**, 2, 1-24.

Bramley, T. & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Bramley, T. & Gill, T. (2008). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, **25**, 3, 293-317.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**, 1, 101-129.

Crisp, V. (2008a). Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education*, **38**, 2, 247-264.

Crisp, V. (2008b). Do assessors pay attention to appropriate features of student work when making assessment judgments? *Research Matters: A Cambridge Assessment Publication*, **6**, 5-9.

Edwards, E., & Adams, R. (2002). *A comparability study in GCE AS geography including parts of the Scottish Higher grade examination. A study based on the summer 2001 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: verbal reports as data*. London: MIT Press.

Fearnley, A. (2000). *A comparability study in GCSE mathematics. A study based on the summer 1998 examination*. Organised by the Assessment and Qualifications Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.

Gill, T., & Bramley, T. (2008). *How accurate are examiners' judgments of script quality? An investigation of absolute and relative judgments in two units, one with a wide and one with a narrow 'zone of uncertainty'*. Paper presented at the British Educational Research Association annual conference, Heriot-Watt University, Edinburgh, September 2008.

Greatorex, J., & Nadas, R. (2008). *Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions?* Paper presented at the British Educational Research Association annual conference, Edinburgh, September 2008.

Jones, B., Meadows, M., & Al-Bayatti, M. (2004). *Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003*. Assessment and Qualifications Alliance.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., & Bay, L. (2001). Setting performance standards using the Body of Work method. In: G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives*. 219-248. Mahwah, NJ: Lawrence Erlbaum Associates.

Leighton, J.P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, **23**, 4, 6–15.

Linacre, J.M. (1992). Treatment Effects. *Rasch Measurement Transactions*, **6**, 2, 218-219. Available at http://www.rasch.org/rmt/rmt62b.htm (Accessed 17/09/09).

Linacre, J. M. (2005). FACETS Rasch measurement computer program. Chicago: Winsteps.com

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, **84**, 3, 231-259.

Ofqual (2009). *GCSE, GCE and AEA Code of Practice*. London: Qualifications and Curriculum Authority. Retrieved April 28, 2009, from http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf.

Rumm Laboratory Pty Ltd. (2004). *Interpreting RUMM2020. Part 1: dichotomous data*. Retrieved June 16, 2008, from http://www.rummlab.com.au/demo.html.

Scharaschkin, A., & Baird, J.-A. (2000). The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343-357.

Suto, W. M. I. & Greatorex, J. (2008) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 2, 213 – 233.

Suto, W.M.I. & Novaković, N. (*in press*). An exploration of the script features that most influence expert judgements in three methods of determining examination grade boundaries. *Assessment in Education: Principles, Policy & Practice*.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273-286.

Appendix A – Instructions to judges

You have 20 separate packs each containing 4 scripts. Each script is identified by a two-letter code, e.g. 'CF'. The label on each pack lists the IDs of the scripts it contains. The scripts you receive within each pack are in no particular order, and have been cleaned of marks. Each judge has a different combination of scripts in their packs.

The task we would like you to complete for each pack is to place the 4 scripts into a single rank order from best to worst.

Although this may not be easy in the absence of marks, we hope that your familiarisation with the question papers, mark schemes, item level statistics and other information from the award meeting will make it a feasible task.

You may use any method you wish to create your ranking, based on scanning the scripts and using your own judgment to summarise their relative merits, but you should not re-mark the scripts. We are expecting each pack to take around 15 minutes to rank, but would also expect the first few packs to take a bit longer while you become accustomed to the task.

No tied ranks are allowed. If you are concerned that two or more scripts are genuinely of exactly the same standard you may indicate this by placing a bracket around them in the table on the record sheet, but you must enter every script onto a separate line of the table, as in the example below:

	Rank	*Script ID*
Best	1	AF
↑	2	DM
↓	3	RO
Worst	4	WP

When you have finished ranking a pack, please replace the scripts in the plastic wallet and return it to the box at the front.

In most packs, the scripts cover a range of about 5-6 marks. Occasionally the range is narrower or wider than this.

Pack 1 contains scripts from the top end of the mark range, working down to Pack 10 which contains scripts from the bottom end of the mark range. The mark ranges of consecutive packs overlap.

Packs 11 to 20 follow the same pattern and use the same scripts, but in different pack combinations.

Please do not collaborate or confer with any of your colleagues who are completing this exercise as it is important that we have independent individual responses.

Appendix B – FACETS output

B641 script features project 05-22-2008 11:02:54

Table 7.2.1 Judge Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Judge
60	120	.5	.50	.00	.23	.79	-2.0	.75	-1.0	1.31	3	
60	120	.5	.50	.00	.23	1.04	.4	1.00	.1	.93	4	
60	120	.5	.50	.00	.24	.84	-1.3	.84	-.3	1.22	6	
60	120	.5	.50	.00	.23	.85	-1.3	.64	-1.3	1.28	1	
60	120	.5	.50	.00	.24	1.13	1.0	1.07	.3	.83	2	
60	120	.5	.50	.00	.25	1.13	1.0	1.14	.4	.81	5	
60.0	120.0	.5	.50	.00	.24	.96	-.4	.91	-.3			Mean (Count: 6)
.0	.0	.0	.00	.00	.01	.14	1.2	.18	.7			S.D. (Populn)
.0	.0	.0	.00	.00	.01	.15	1.4	.19	.8			S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability .83
Model, Fixed (all same) chi-square: .0 d.f.: 5 significance (probability): 1.00

Table 7.3.1 Script Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Script
9	18	.5	1.00	8.87	.77	.78	-.2	.47	-.4	1.20	42	F01B (mark 50)
9	18	.5	1.00	8.33	.68	1.12	.4	.91	.2	.88	21	E01A (mark 48)
9	18	.5	1.00	8.08	.58	1.28	1.0	2.69	2.2	.25	12	C06B (mark 49)
9	18	.5	1.00	7.39	.73	1.58	1.2	1.96	1.0	.39	10	C05B (mark 47)
9	18	.5	1.00	7.15	.59	.58	-1.6	.43	-1.0	1.69	9	C05A (mark 47)
9	18	.5	1.00	7.13	.61	.65	-1.0	.56	-.6	1.45	71	Z06A (mark 52)
9	18	.5	1.00	6.89	.57	.85	-.5	.69	-.6	1.31	22	E01B (mark 48)
9	18	.5	1.00	6.82	.53	1.13	.6	1.18	.7	.65	41	F01A (mark 50)
9	18	.5	1.00	6.35	.58	.70	-1.0	.76	-.5	1.40	72	Z06B (mark 52)
9	18	.5	1.00	6.13	.58	.77	-.8	.60	-.8	1.42	11	C06A (mark 49)
9	18	.5	1.00	5.59	.65	1.23	.6	.93	.0	.84	62	Z01B (mark 46)
9	18	.5	1.00	5.41	.66	.78	-.4	.51	-.4	1.29	26	E03B (mark 40)
9	18	.5	.99	5.22	.59	.46	-2.1	.36	-1.4	1.79	1	C01A (mark 44)
9	18	.5	.99	5.06	.60	1.22	.7	1.39	.8	.65	51	F06A (mark 45)
9	18	.5	.99	4.57	.60	1.35	1.2	2.65	1.9	.21	24	E02B (mark 43)
9	18	.5	.99	4.51	.62	1.13	.5	1.47	.7	.73	61	Z01A (mark 46)
9	18	.5	.99	4.36	.59	1.30	1.0	1.11	.3	.56	2	C01B (mark 44)
9	18	.5	.99	4.33	.57	.79	-.8	.59	-.6	1.47	52	F06B (mark 45)
9	18	.5	.98	3.95	.57	.92	-.2	1.34	.8	.99	43	F02A (mark 41)
9	18	.5	.98	3.95	.61	.97	.0	.66	-.2	1.14	54	F07B (mark 35)
9	18	.5	.98	3.76	.63	.94	-.1	.57	.2	1.21	30	E05B (mark 35)
9	18	.5	.97	3.38	.58	.91	-.2	.76	-.4	1.19	25	E03A (mark 40)
9	18	.5	.96	3.30	.54	1.05	.3	.90	.0	.94	27	E04A (mark 38)
9	18	.5	.96	3.30	.58	.58	-1.6	.44	-1.0	1.70	13	C07A (mark 41)
9	18	.5	.96	3.28	.62	1.08	.4	.83	.4	.89	14	C07B (mark 41)
9	18	.5	.96	3.09	.59	1.21	.8	1.44	.7	.55	44	F02B (mark 41)
9	18	.5	.95	3.03	.60	.82	-.5	.60	.0	1.33	28	E04B (mark 38)
9	18	.5	.94	2.79	.55	1.04	.2	.95	.1	.92	23	E02A (mark 43)
9	18	.5	.90	2.14	.57	.95	.0	.93	.0	1.08	73	Z07A (mark 36)
9	18	.5	.89	2.13	.62	1.01	.1	.75	.0	1.05	64	Z02B (mark 39)
9	18	.5	.86	1.78	.61	.74	-.7	.56	-.5	1.39	63	Z02A (mark 39)
9	18	.5	.80	1.41	.60	.74	-.8	.66	-.2	1.39	46	F03B (mark 34)
9	18	.5	.78	1.24	.63	1.08	.3	1.35	.6	.82	53	F07A (mark 35)
9	18	.5	.74	1.05	.60	.71	-1.1	.50	-.4	1.52	29	E05A (mark 35)
9	18	.5	.64	.58	.54	.79	-.9	.73	-.7	1.52	31	E06A (mark 30)
9	18	.5	.64	.56	.75	1.68	1.4	1.29	.6	.48	74	Z07B (mark 36)
9	18	.5	.61	.45	.60	.99	.0	.81	.0	1.05	16	C08B (mark 31)
9	18	.5	.59	.36	.61	1.21	.8	1.79	.9	.40	4	C02B (mark 33)
9	18	.5	.58	.32	.64	.86	-.4	.52	-.2	1.29	56	F08B (mark 27)
9	18	.5	.50	.01	.67	.75	-.7	.42	-.6	1.36	45	F03A (mark 34)
9	18	.5	.43	-.27	.57	1.52	2.1	1.90	.9	-.40	66	Z03B (mark 28)
9	18	.5	.41	-.35	.59	.69	-1.3	.48	-.1	1.62	15	C08A (mark 31)
9	18	.5	.41	-.37	.67	.73	-.6	.45	-.4	1.34	3	C02A (mark 33)
9	18	.5	.41	-.38	.56	.65	-1.5	.53	-1.1	1.73	32	E06B (mark 30)
9	18	.5	.35	-.62	.56	.89	-.4	.73	.0	1.31	55	F08A (mark 27)
9	18	.5	.26	-1.04	.89	.38	-1.1	.15	-.5	1.40	5	C03A (mark 24)
9	18	.5	.22	-1.26	.77	1.42	.9	1.17	.5	.71	77	Z09A (mark 25)
9	18	.5	.10	-2.19	.71	.47	-1.2	.29	-1.0	1.46	65	Z03A (mark 28)
9	18	.5	.07	-2.65	.76	.67	-.5	.45	-.5	1.26	6	C03B (mark 24)
9	18	.5	.06	-2.71	1.00	1.59	.9	1.43	.8	.71	75	Z08A (mark 32)
9	18	.5	.05	-2.96	1.02	1.55	.8	1.01	.6	.77	76	Z08B (mark 32)
9	18	.5	.05	-2.98	.63	1.24	.7	1.53	1.0	.67	58	F09B (mark 19)
9	18	.5	.04	-3.09	.69	.70	-.6	.52	-.3	1.30	39	E10A (mark 18)
9	18	.5	.03	-3.33	.65	.42	-1.8	.29	-1.5	1.61	18	C09B (mark 25)
9	18	.5	.03	-3.36	.62	.75	-.6	.75	-.4	1.28	36	E08B (mark 23)
9	18	.5	.02	-3.85	.70	1.55	1.2	1.49	.7	.44	34	E07B (mark 26)
9	18	.5	.02	-3.86	.65	1.42	1.2	1.43	.7	.45	33	E07A (mark 26)
9	18	.5	.02	-3.98	.56	.86	-.4	.83	-.3	1.25	19	C10A (mark 16)
9	18	.5	.02	-4.17	.62	.97	.0	.76	-.1	1.08	37	E09A (mark 20)
9	18	.5	.01	-4.34	.57	1.00	.0	.89	.0	1.04	40	E10B (mark 18)
9	18	.5	.01	-4.38	.60	.76	-.7	.58	-.8	1.39	60	F10B (mark 13)
9	18	.5	.01	-4.51	.57	.79	-1.0	.59	-.5	1.55	59	F10A (mark 13)
9	18	.5	.01	-4.56	.60	.48	-1.9	.38	-1.2	1.73	38	E09B (mark 20)
9	18	.5	.01	-4.68	.55	.88	-.4	.73	-.3	1.31	68	Z04B (mark 21)
9	18	.5	.01	-4.73	.54	.88	-.5	.83	-.2	1.31	20	C10B (mark 16)
9	18	.5	.01	-4.80	.59	1.06	.3	1.07	.3	.89	80	Z10B (mark 17)
9	18	.5	.01	-4.84	.54	1.33	1.6	1.40	.9	.04	49	F05A (mark 14)
9	18	.5	.01	-5.05	.59	1.05	.2	.85	.1	.98	78	Z09B (mark 25)
9	18	.5	.01	-5.21	.54	.90	-.4	.77	-.2	1.31	35	E08A (mark 23)
9	18	.5	.00	-5.38	.61	1.17	.6	1.11	.3	.77	48	F04B (mark 22)
9	18	.5	.00	-5.46	.57	.91	-.3	.78	-.1	1.20	67	Z04A (mark 21)
9	18	.5	.00	-5.57	.59	1.66	2.0	2.24	2.1	-.18	7	C04A (mark 14)
9	18	.5	.00	-5.66	.55	.93	-.2	.78	-.4	1.21	79	Z10A (mark 17)
9	18	.5	.00	-5.83	.62	1.02	.1	.88	.1	.98	47	F04A (mark 22)
9	18	.5	.00	-6.15	.82	1.24	.6	.85	.6	.79	17	C09A (mark 25)
9	18	.5	.00	-6.27	.58	.72	-.9	.55	-1.0	1.48	50	F05B (mark 14)
9	18	.5	.00	-6.29	.57	.72	-1.0	.56	-.9	1.51	57	F09A (mark 19)
9	18	.5	.00	-6.36	.57	1.06	.2	1.00	.1	.92	69	Z05A (mark 12)
9	18	.5	.00	-6.67	.60	1.06	.2	.94	.0	.94	8	C04B (mark 14)
9	18	.5	.00	-7.89	1.04	.99	.2	.62	.0	1.04	70	Z05B (mark 12)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu	Script
9.0	18.0	.5	.49	.00	.63	.97	-.1	.91	.0			Mean (Count: 80)
.0	.0	.0	.44	4.54	.10	.30	.9	.50	.8			S.D. (Populn)
.0	.0	.0	.44	4.56	.10	.30	.9	.50	.8			S.D. (Sample)

Model, Populn: RMSE .64 Adj (True) S.D. 4.49 Separation 7.03 Reliability .98
Model, Sample: RMSE .64 Adj (True) S.D. 4.52 Separation 7.07 Reliability .98
Model, Fixed (all same) chi-square: 4352.6 d.f.: 79 significance (probability): .00
Model, Random (normal) chi-square: 78.1 d.f.: 78 significance (probability): .48

Starting them young: research and project management opportunities for 16 to 19 year olds

Irenka Suto and Rita Nádas Research Division

Introduction

Several educational routes have been developed which entail project work with a specific focus on independent learning and research. In this article, we outline some of the options that exist at Level 3, primarily for 16 to 19 year olds (Years 12 and 13) in the UK and internationally. We then conduct a more detailed comparison of two routes: the Extended Project Qualification, and the International Baccalaureate Extended Essay. Many stakeholders may be unaware of the differences in the aims, structure, and scope of these routes. It is important for students and teachers to be conscious of the differences so that they can make informed decisions about what is most suitable for them. End-users such as higher education admissions tutors and employers also need to understand the differences in order to weigh up the experiences and achievements of applicants fairly.

Research and project routes for 16 to 19 year olds

As is the case in many countries, students aged 16 to 19 in the UK are able to choose which subjects they study, and whether the educational route followed is general or vocational in nature, or a combination of the two. Some students carry out independent research and investigative work as part of GCE Advanced (A) level courses, although opportunities vary among subjects. In a recent study of A level teaching, for example, Mehta, Suto, Elliott and Rushton (2011) report that half of A level French and Economics teachers set their students investigation and/or research tasks. In contrast, most A level Mathematics teachers felt that much of their course does not require an independent approach on the part of the students. Opportunities for investigative and research tasks within mainstream vocational courses are also likely to vary. In contrast, other educational routes have been designed solely to offer students the chance to conduct an independent project. These opportunities are often more substantial and specific than those embedded within subject-based courses. Some major examples are outlined below.

Farnborough extended projects

The Sixth Form College, Farnborough, a large state-funded college in the south of England, has run its own scheme of 'extended projects' since 2006. The scheme was devised in the wake of the Tomlinson report (2004) which alleged a lack of opportunity for students to practise effortful study; that is, to engage in deep learning rather than absorbing inert knowledge. Farnborough students are encouraged to go beyond the comfort zones of their A level course material and to investigate any topic of particular interest to them that links at least two of their A level subjects. Students conduct their projects and write up formally structured 5000-word reports between May of Year 12 and October of Year 13, utilising their free time during their summer holidays.

Farnborough students carry out their projects with the support of an assigned supervisor who is also a teacher at the college. Assessment is internal, taking the form of detailed written comments which use Dweck's (1999) 'two stars and a wish' model to provide formative and instructive feedback. The two stars relate to two points of praise which focus on the project task (rather than on the student him/herself). The wish is a suggestion as to how the work could be extended, taken forward, and improved. Staff at the college have chosen to avoid summative assessment and formal accreditation of the extended projects as qualifications in the belief that this would restrict risk-taking behaviour. Arguably, assigning numerical or alphabetical values to project reports encourages students to play safe and focus on assessment criteria, rather than following their academic interests and instincts.

Extended Project Qualification (EPQ)

A scheme of nationally available project qualifications has evolved from the scheme devised by The Sixth Form College, Farnborough. Project qualifications are now administered by at least five major awarding bodies in the UK (OCR, AQA, Edexcel, Education Development International plc [EDI], the Welsh Joint Education Committee [WJEC], and VTCT[1]). They are an option for secondary school students at three different levels: Foundation (Level 1), Higher (Level 2), and Extended (Level 3) (Ofqual, 2011). At Level 3, Extended Project Qualifications (EPQs) can be taken alongside A levels, as well as being a compulsory element in Diplomas. (For further information on Diplomas, see Ertl, Stanley, Huddleston, Stasz, Laczik and Hayward, 2009.) Like Farnborough students, EPQ students explore a personal interest, engaging in cross-curricular study which may take place both inside and outside the classroom. In contrast, however, the topic of the EPQ does not have to be related to anything else that the student is studying, and its outcome does not have to be a traditionally styled piece of academic scholarship. Although many EPQs culminate in a 5000-word dissertation, it is equally acceptable for students to produce a shorter report accompanying an artefact such as a piece of art, furniture, or music.

EPQs are assessed summatively; originally, staff at The Sixth Form College, Farnborough worked closely with the AQA awarding body to develop an initial mark scheme for assessing the projects. Other versions of the mark scheme, grounded in generic descriptive assessment criteria, have since been developed by other awarding bodies. EPQs are intended to engender so-called 21st Century skills such as creativity and imagination, problem-solving skills, independent thinking, cooperation with others, and using people as resources (Department for Education and Skills, 2005), and these skills are the focus of assessment. Students are

1 VTCT is a specialist awarding body offering vocational qualifications. The acronym is not explained on its website: http://www.vtct.org.uk/

rewarded for the process undertaken, rather than for the quality of the outcome of their endeavours.

As can be seen in Table 1, which contains data from the National Pupil Database, the popularity of EPQs has increased rapidly since they first became available nationally in 2007. Most recently, in the summer of 2011, the Joint Qualifications Council reported a further 51% rise in the number of grades issued for EPQs. It should be noted, however, that absolute numbers are small, and the research route is still followed by only a minority of Year 13 students in England.

Table 1: Uptake of the Extended Project Qualification (EPQ)*

Examination session	*EPQ Year 13 candidates in England*	*All Year 13 candidates in England*	*Percentage of all Year 13 candidates in England taking EPQ in the summer exam session*
Summer 2007	17	323,688	0.01%
Summer 2008	919	339,468	0.27%
Summer 2009	3350	365,717	0.92%
Summer 2010	11492	392,176	2.93%

*Data obtained from the National Pupil Database (Department for Education)

Cambridge Pre-U Independent Research Report (IRR)

Another rapidly expanding research route is the Independent Research Report (IRR), a major component of the Cambridge Pre-U Certificate in Global Perspectives and Research (GPR). GPR was developed and is administered by Cambridge International Examinations (CIE), who publish UK and international editions of the syllabus. According to CIE (2011), GPR is taught as two successive one-year courses: in Year 13, the IRR grows seamlessly out of the skills introduced and developed in a Global Perspectives course, which is taught in Year 12. The IRR takes forward the Year 12 Global Perspectives course's emphasis on an interdisciplinary, independent and reflective approach to education, focusing on the need for rigour in the analysis and construction of arguments (CIE, 2008). Its explicit aims are to: prepare students for a way of working in higher education; develop generic and higher order skills of research and analysis; and encourage intellectual curiosity (*ibid*).

IRR students submit a report based on work they have done on self-chosen topics beyond individual subject syllabuses (e.g. A levels or Cambridge Pre-U Principal Subjects). They may choose to: (i) dig deeper in a chosen specialism, (ii) cross subject boundaries with an inter–disciplinary enquiry, or (iii) make a new departure with a study in a non-school subject such as astronomy or anthropology (CIE, 2008). The report must be a single piece of extended writing in the form of a 4500 to 5000-word dissertation or report based on an investigation or field study (*ibid*). Assessment is summative. It focuses on abilities to: design, plan and manage a research project; collect and analyse information; evaluate and make reasoned judgements; and communicate findings and conclusions (*ibid*).

IB Extended Essay

A further research route available to students internationally is the Extended Essay undertaken by students of the International Baccalaureate Organisation's Diploma programme (International Baccalaureate Organisation [IBO], 2011). IB students engage in independent research through an in-depth study of a question relating to one of the subjects they are studying. They write essays of up to 4000 words which are marked summatively and externally (by teachers from other IB schools and colleges). Short concluding interviews are also held with students' supervisors. The Extended Essay is intended to promote 'high-level research and writing skills, intellectual discovery and creativity' (IBO, 2011). As with the IRR, it enables students to practise the thesis approach to writing that is subsequently needed at many universities, whilst experiencing the excitement of intellectual challenge and discovery. Further details are considered subsequently.

Other research routes

In addition to the better-known routes described previously, award schemes run by national associations and funding bodies provide subject-specific opportunities for 16 to 19 year olds to carry out research and project work. One example is the British Science Association's (2011) scheme of *Crest* awards. The scheme operates at three different levels of secondary education (from 11 to 19 years), and students are rewarded for undertaking individual or team-based project work in science, technology, engineering and mathematics (STEM subjects). At the highest of the three levels, over seventy hours of work are put into projects. The project work may link both into the school or college curriculum and into work experience placements and after-school clubs. Another example is a project undertaken by A level science students at Simon Langton Grammar School for Boys in Canterbury. A *People* award from the Wellcome Trust was used to support collaborative research between the school and the University of Kent. The project entailed students using basic genetic engineering techniques in experiments conducted during lunchtimes and free periods. Ultimately, it fed into research to help to understand the causes of multiple sclerosis (Wellcome Trust, 2008.)

Comparison of the EPQ and the IB Extended Essay

The EPQ and the IB Extended Essay are two of the most widely followed research routes. We compared the OCR specification for the EPQ (OCR, 2011) with documentation published to support the Extended Essay (IBO, 2004, 2007). Several different dimensions were considered in the comparison, relating to the two research routes' structures, skills focuses, and assessment approaches.

The key structural features are summarised in Table 2. It can be seen that the EPQ requires three times as great a time commitment as does the Extended Essay, from both students and supervisors, although it is unknown how much time is actually spent. EPQ outcomes can be comparatively more varied in structure, format, and the topic covered. The Extended Essay, on the other hand, is always linked closely with students' other studies, and its format is prescribed more tightly. Extended Essays may therefore be less diverse, but a more consistent entity for end-users to evaluate.

Table 3 contains details of the knowledge, skills, and understanding that EPQ students and Extended Essay students aim to acquire. The exact wording used in documentation associated with the two research routes (IBO, 2004, 2007; OCR, 2011) is used wherever possible. Some of the skills are presented as so-called '21st Century' skills, and have been defined and grouped in line with the work of a major international collaboration: *Assessment and Teaching of 21st Century Skills* (ATC21S, 2011). Other types of knowledge, skills and understanding are presented lower in the table. It is evident that there is some overlap between the EPQ and Extended Essay: both types of students aim to acquire skills in creativity, critical thinking, communication, research, and personal responsibility. The EPQ differs from the Extended Essay, however, in that greater emphasis is placed on project management, and there is less explicit emphasis on in-depth knowledge and understanding, and on intellectual risk-taking and discovery.

Table 2: Key structural features of the EPQ and the IB Extended Essay

Feature	*OCR EPQ*	*IB Extended Essay*
Positioning	Can be a stand-alone linear qualification worth half an A level (20 to 70 UCAS tariff points*), or a component of the Level 3 Diploma.	Compulsory component of the IB Diploma.
Format of student outcome	5000-word dissertation, or other outcome (design, artefact, report, performance) accompanied by a 1500 to 2500-word report. All students must also complete a Project Progression Record, which contains details of all activities undertaken, and the supervisor's comments on them.	<4000-word essay
Topic choice	The student can choose any topic with agreement from his or her supervisor. If the project is part of the Diploma, then the topic should be linked, as appropriate, to the Principal Learning, that is, the subject area being studied.	The student must choose a topic that fits into one of the subjects on the approved Extended Essay list. It is normally within one of the student's six chosen subjects for the IB diploma. The student chooses the topic in cooperation with his or her supervisor.
Language	Always written in English.	Essays on literary topics are written in the student's 'mother tongue'. Students studying a second modern language (e.g. Japanese) may write an essay in this target language, in which case the research topic must be related to the target culture. All other essays (e.g. on scientific topics) are written in English, French or Spanish.
Structure of written outcome	Not specified.	The formal requirements of the final outcome are: • Title page • Abstract • Contents page • Introduction • Body (development/methods/ results) • Conclusion • References and bibliography • Appendices
Recommended time requirements	120 Guided Learning Hours (50 hours linked to teaching & 70 hours linked to assessment). NB: this excludes self-directed study time.	Supervisors should spend between 3 and 5 hours with each student, including the time spent on a viva voce. Students should work for 40 hours on their essays.
Main document providing an overview of the research route	*The Extended Project Level 3 handbook/specification*. Includes sections on: • Learning outcomes • Assessment objectives • Assessment criteria, with exemplifications of what the learner will do • Marking criteria • Glossary (2 pages of terms used in the assessment/ marking grids) • Information on key skills, functional skills, and personal learning and thinking skills	*Diploma Programme Extended Essay Guide*. Includes sections on: • IB learner profile • Aims • Assessment objectives • Details – all essays (including assessment criteria [~OCR EPQ 'marking criteria']) • Details – subject-specific (c. 5 pages on each of 27 subjects, including subject-specific interpretations of assessment criteria)

*UCAS is the Universities and Colleges Admissions Service. For information about UCAS tariff points, see http://www.ucas.com/students/ucas_tariff/

Table 3: Knowledge, skills, and understanding students are aiming to acquire in the EPQ and the IB Extended Essay

21st Century skills mentioned* explicitly *in aims and in statements of intention and opportunities*	*OCR EPQ*	*IB Extended Essay*
1. Creativity and innovation	Creativity Innovation (initiative and enterprise)	Creativity
2. Critical thinking, problem-solving, decision-making	Critical thinking Problem-solving Decision-making	Critical thinking (including constructing reasoned arguments)
3. Learning to learn, metacognition	Develop and improve own learning and performance	Not mentioned
4. Communication	Communication (including presentation skills)	Communication (including high level writing skills)
5. Collaboration (teamwork)	Project may be a defined task within a collaborative group project	Not mentioned
6. Information literacy (includes research on sources, evidence, biases, etc.)	Understand and use research skills	Research (including a concern with interpreting and evaluating evidence)
7. ICT literacy	Where appropriate, e-confidence - applying new technologies	Not mentioned
8. Citizenship – local and global	Not mentioned	Not mentioned
9. Life and career	Use learning experiences to support personal aspirations for further study and/or career development	Not mentioned
10. Personal and social responsibility – including cultural awareness and competence	Responsibility either for an individual task or for a defined task within a group project	Personal responsibility for own independent learning
Other knowledge, skills, and understanding also mentioned in aims and in statements of intention and opportunities	• Project management • Design, planning, research, analysis, synthesis, and evaluation • Development as critical, reflective and independent learners • 'Learners will have the opportunity to apply and develop their personal learning and thinking skills (PLTS), the functional skills of English, mathematics and information and communication technology (ICT) and key skills'	• Engagement in a systematic process of independent research appropriate to the subject • Excitement of intellectual discovery • Intellectual risk-taking and reflection • Open-mindedness, balance and fairness • In-depth knowledge and understanding

*as defined and grouped by a major international collaboration: *Assessment and Teaching of 21st Century Skills* (ATC21S); see http://atc21s.org.

Key aspects of the assessment approaches of the two research routes are collated in Table 4. Perhaps the most striking difference between them is that EPQ assessment focuses *exclusively* on the process of undertaking a project. In contrast, Extended Essay assessment takes account of the outcome of the research process, as well as the process *per se*. EPQ assessment is internal, emphasising evaluation, review, and critical work, whereas the Extended Essay assessment emphasises argument and analysis and focuses to a greater extent on academic writing skills.

Table 4: Assessment approaches of the EPQ and the IB Extended Essay

Aspect of assessment	*OCR EPQ*	*IB Extended Essay*
Assessors	Following within-centre standardisation (in centres with multiple entries), the project is marked by the student's own supervisor. A sample of work from each school/college is moderated externally by OCR-appointed moderators.	Completion of the written essay should be followed by a short, concluding interview, or *viva voce* (10-15 minutes), with the supervisor. The essay is marked externally by IBO-appointed examiners.
Marks awarded	A total score is obtained on a scale of 0 to 60. The stand-alone qualification is graded as A*–E. For Diploma students, the grade is translated into a points score. The overall Diploma grade is calculated by adding this to the point score for the student's Principal Learning.	A total score is obtained on a scale of 0 to 36. This is used to determine the band (A, B, C, D, or E) in which the essay is placed. This band, in conjunction with the band awarded for the Theory Of Knowledge (TOK) component, determines the number of IB Diploma points (0 to 3) awarded for these two requirements.
Focus (content/process)	The focus of the assessment is on the process the learner has gone through to achieve and evaluate their final outcome rather than the outcome itself.	Emphasis is placed on the research process and its formal outcomes.
Objectives	• Manage • Use resources • Develop & realise • Review & communicate	• Plan & pursue a research project • Formulate research question • Gather & interpret • Structure argument • Present • Use terminology & language • Apply analytical & evaluative skills
Assessment terms mentioned most frequently in main documentation	• *Evaluation*, *review*, and *critical*	• *Argument* and *analysis*
Judgement of student work against assessment criteria	Best fit approach entailing holistic judgement	Best fit 'bottom up' approach entailing holistic judgement

Discussion

In recent years, research and project management has become an important component of education for many 16 to 19 year olds. The increasing interest is unsurprising, given the high levels of competition for jobs and university places that currently exist, which place pressure on applicants to distinguish themselves from their competitors. Key questions for educationalists and policy-makers relate to whether, and the extent to which, research projects conducted outside of mainstream vocational and general courses such as A levels should be influencing university admissions tutors and employers. These are important concerns, given the diverse opportunities and resources available to different student groups nationally and internationally. Although project grades (where they exist) are not always part of formal conditional offers for university places, which tend to relate to A levels and similar courses, research projects can be described in application forms and discussed in interviews. In this way, some applicants can use their project work to demonstrate their commitment and enthusiasm for a particular subject or for education in general. Whilst problems relating to inequality of opportunities affect an educational context far broader than the one considered here, the variation in research and project work among 16 to 19 year olds could be considered an important example of how the problem shows no signs of abating.

The differences among research routes highlighted in this article indicate the breadth of skills being nurtured in young people. Positive experiences of all routes are easy to find, and no single one can be said to meet all needs. Whilst some students may be striving to develop specialist academic research skills and deepen their subject knowledge beyond A level, others may be focused on acquiring the generic skills and capabilities considered most desirable by many employers (Confederation of British Industry, 2007). It is important that admissions tutors, employers, and other end-users understand some of the differences in approach, focus and project magnitude, so that they can evaluate and compare applicants' accomplishments meaningfully. Furthermore, this will aid them in clarifying what they prefer future applicants, teachers, and course developers to devote time to. Stakeholder engagement of this kind is critical in ensuring that young people are sufficiently prepared for the challenges of higher education and professional life.

Finally, it is worth reflecting on the value of conducting summative assessments of student projects. The variety of assessment provision described in this article reflects the diversity of views on this issue held among highly experienced educationalists. On the one hand, because the EPQ and IB Extended Essay are assessed summatively, they can constitute or contribute to formal qualifications. Accreditation of the EPQ by the national regulator (Ofqual, 2011) has ensured that schools and colleges across the country have the financial means to offer it to their students. Furthermore, grades may provide extrinsic motivation for some students, and teachers who lack the skills or experiences needed to set up their own project schemes can obtain support and advice from awarding bodies. On the other hand, the case for encouraging academic risk-taking through formative assessment, or even non-assessment, is also powerful. An appreciation of the intrinsic value and intellectual satisfaction of undertaking a project is arguably something to be nurtured more actively among young people.

References

ATC21S (2011). http://atc21s.org Accessed 06/10/11.

British Science Association (2011). http://www.britishscienceassociation.org/web/ccaf/CREST/index.htm Accessed 02/06/11.

Cambridge International Examinations (2008). *Cambridge International Level 3 Pre-U Certificate in Global Perspectives and independent research. Cambridge Pre-U Syllabus – International edition*. Cambridge: University of Cambridge Local Examinations Syndicate.

Cambridge International Examinations (2011). http://www.cie.org.uk/qualifications/academic/uppersec/preu/subjects/subject/preusubject/?assdef_id=1018

Confederation of British Industry (2007). *Shaping up for the future: The business vision for education and skills*. www.cbi.org.uk/bookshop Accessed 06/04/11.

Department for Education and Skills (2005). *The 14–19 Education and Skills White Paper*. London: HMSO.

Dweck, C.S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Taylor and Francis/Psychology Press.

Ertl, H., Stanley, J., Huddleston, P, Stasz, C., Laczik, A. & Hayward, G. (2009). *Reviewing Diploma Development: Evaluation of the Design of the Diploma Qualifications*. Research Report No. DCSF-RW080. Oxford: Oxford University Press.

Joint Council for Qualifications (2011). News release for A, AS and AEA results, summer 2011. http://www.jcq.org.uk/attachments/published/1584/JCQ%20Results%20Press%20Notice%2018%20August%202011.pdf Published 18/08/11.

International Baccalaureate Organisation (2004). *Diploma Programme assessment: Principles and practice*. Cardiff, Wales: International Baccalaureate Organisation.

International Baccalaureate Organisation. (2007). *Diploma programme extended essay - guide (for examinations in 2009)*. Cardiff, Wales: International Baccalaureate Organisation.

International Baccalaureate Organisation (2011). http://www.ibo.org/diploma/curriculum/core/ Accessed 03/06/11.

Mehta, S., Suto, I., Elliott, G., & Rushton, N. (2011). *Independent Research at A level: Students' and Teachers' Experiences*. Presented at the annual conference of the British Educational Research Association, 6-9 September, 2011.

OCR (2011). *Level 3 Extended Project: Centre handbook/specification*.

Ofqual (2011). http://www.ofqual.gov.uk/qualifications-assessments Accessed 03/10/11.

Tomlinson, M. (2004). *14-19 Curriculum and Qualifications Reform: Final Report of the Working Group on 14-19 Reform*.

Wellcome Trust, (2008). http://www.wellcome.ac.uk/News/2008/News/WTX052388.htm Accessed 04/10/11.

An investigation into the impact of screen design on computer-based assessments

Matt Haigh Research Division

Introduction

Many authors put validity at the heart of assessment (Kane, 2006; Popham, 2000) and emphasise the importance of validity in evaluating new forms of assessment. For example,

> *The arguments… regarding traditional and alternative forms of assessment need to give primacy to evolving conceptions of validity if, in the long run, they are to contribute to the fundamental purpose of measurement - the improvement of instruction and learning.* (Linn *et al.*, 1991, p.20)

Arguments have also been put forward demonstrating the role of computer-based assessment (CBA) in both enhancing and reducing the validity of test scores. Ridgway and McCusker (2003) highlight benefits of CBA in improving the validity of assessing problem-solving skills, whilst Clarke *et al*. (2000) identify the detriment to validity from dependence on multiple choice items. Throughout the history of CBA, there has been discussion regarding the validity aspects of its implementation (Huff and Sireci, 2001; Russell, Goldberg, and O'Connor, 2003).

Educational measurement theory emphasises construct validity in evaluating test outcomes (Messick, 1989). Construct validity is defined as "the qualities a test measures, determined by the degree to which certain explanatory concepts or constructs account for performance on the test" (Messick, 1989, p.16).

Construct validity can be affected by 'construct-irrelevant variance'; it occurs when the test contains excess variance that is irrelevant to the interpreted construct. For example, a demanding reading stimulus in a science assessment may cause a variance in test scores (related to reading ability) that is irrelevant to the construct being assessed (science).

Some aspects of construct-irrelevant variance have been explored in the CBA literature. A number of studies indicate that students with a good prior knowledge of ICT performed better on computer-based tests (Clariana and Wallace, 2002; Russell *et al*., 2003; Warschauer, 2004). Construct-irrelevant variance can be introduced by poor item design (McKenna, 2001; Sireci and Zenisky, 2006); screen size and resolution (Bridgeman, Lennon, and Jackenthal, 2003); and the effect of scrolling (Ricketts and Wilks, 2002). These studies indicate that aspects of the screen environment or the method of student interaction may be related to sources of construct-irrelevant variance in CBA. Additional research has investigated how the layout of paper-based formats may affect item performance (Crisp and Sweiry, 2006) and how screen design affects how website users access information (Helander, Landauer, and Prabhu, 1997). However, there is no research on how item format[1] may affect performance by students on a computer-based test. This article reports on part of a study that investigated the impact of item format on the difficulty of test items. The following research question was investigated:

> *What are the effects of changing the item format on measures of item difficulty of a computer-based test item?*

1 'Item format' is the term used in this article to cover the layout of text, buttons and images on the computer screen, along with the method of interaction used with these screen elements.

Method

The research question implied a causal relationship between item format and item difficulty, which required a quantitative experimental methodology. Within this paradigm, a 'post-test/observation only with control group' experimental design (Black, 1999) was used.

Two parallel forms of a computer-based test were developed; each test consisted of 15 items based on the GCSE Science curriculum. Five items were identical in both forms of the test to act as a control. The remaining items, shown in the Appendix, were modified in the parallel forms to investigate the effect of the following aspects of item format:

- Presence or absence of colour image.
- Drag and drop categorisation vs. tick-box categorisation.
- Multiple choice single option selection vs. multiple option select.
- Completion by drag and drop vs. drop down selection.
- Matching objects with lines vs. matching objects using a table.
- Static graphic vs. animated graphic.
- Select correct answer vs. drag answer to target.
- Tick-boxes to select statements vs. whole statement selections.
- Visual resources on single page vs. using tabbed panels to move between information.
- Restricted free-text input box vs. unlimited & scrollable free-text input box.

Students from seven secondary schools in England participated in the research; each student was randomly assigned one of the two parallel forms of the test. For each item, two measures of item difficulty were calculated[2]. Each measure was then evaluated for significant differences between the alternate forms of each item. Note that each of the ten aspects of item design were analysed independently.

Findings

Sample

The science test was taken by 112 students and the seven schools varied in size and school type, but were mainly community comprehensives in urban areas. Table 1 shows the background data relating to the sample.

Table 1: Background variables relating to the sample

Measure	*National Mean*	*Form 1 Mean (n=55)*	*Form 2 Mean (n=57)*	*Form 1 Standard Deviation*	*Form 2 Standard Deviation*
National Test Score[3]	18.49	20.16	19.73	2.49	3.21
Predicted GCSE Score[4] Science	4.64	6.42	6.29	1.24	1.31
Total GCSE Point Score[4]	43.20	57.20	55.84	22.10	23.71
ICT competence[5]	n/a	2.22	2.21	0.83	0.98
Score on 5 common items	n/a	7.24	7.67	3.18	3.12

2 Using both Classical Test Theory and Item Response Theory paradigms – see Hambleton, R. K., and Jones (1993) for a useful comparison.

3 Sum of KS2 English level, KS2 Maths level, KS3 English level, KS3 Maths level

4 GCSE score: Grade A*=8 points, Grade A=7 points, B=6, C=5, D=4, E=3, F=2, G=1

5 Self-reported on scale 1= 'Not very good with ICT' to 5 = 'Very competent with ICT'

Measures of student attainment indicated a spread of attainment within the sample, although the mean attainment of the sample was higher than the national mean. Control variables relating to student attainment and ICT competence were not significantly different across the two forms of the test. The mean score on the five identical items was not significantly different in the two forms of the test, indicating the random assignment had produced well-matched samples.

Classical Test Theory - Item Facility Analysis

Item facility is the average number of marks achieved by students for an item expressed as a proportion of the maximum mark. A value of 0 indicates a very difficult item; a value of 1 indicates a very easy item. Table 2 shows the facility values for the items in each of the parallel forms of the test along with outcomes of an independent sample t-test to identify significant differences:

Table 2: Item Facility Measures

Item no.	*Facility Form 1*	*Facility Form 2*	*Difference*	*t-test statistic*	*Significance*
1*	0.43	0.51	-0.08	-0.871	0.386
2*	0.54	0.56	-0.02	-0.375	0.708
3*	0.75	0.75	0.00	-0.003	0.998
4*	0.75	0.75	0.00	-0.003	0.998
5*	0.31	0.36	0.05	-0.580	0.563
6	0.45	0.44	0.01	0.168	0.867
7	0.65	0.63	0.02	0.302	0.763
8	0.27	0.30	-0.03	-0.296	0.767
9	0.82	0.87	-0.05	-0.803	0.424
10	0.60	0.61	-0.01	-0.011	0.992
11	0.52	0.41	0.11	1.254	0.213
12	0.67	0.79	-0.12	-1.394	0.213
13	0.23	0.28	-0.05	-0.674	0.502
14	0.50	0.61	-0.11	-1.109	0.271
15	0.31	0.43	-0.12	-1.416	0.161

*indicates common item

Although differences in difficulty were observed in the parallel forms of each item, the t-test indicates that these were not statistically significant. This suggests that the modifications to item format had very little effect on item-facility in any of the cases. A visual representation of the data is shown by the scatter plot in Figure 1. The numerical labels on the data points correspond to each of the *modified* items (1 = Question 6, 2 = Question 7, 3 = Question 8 etc.) and the diagonal dotted line represents item forms of equal difficulty. The scatter plot shows all items were close to the line of equal difficulty in their alternative forms.

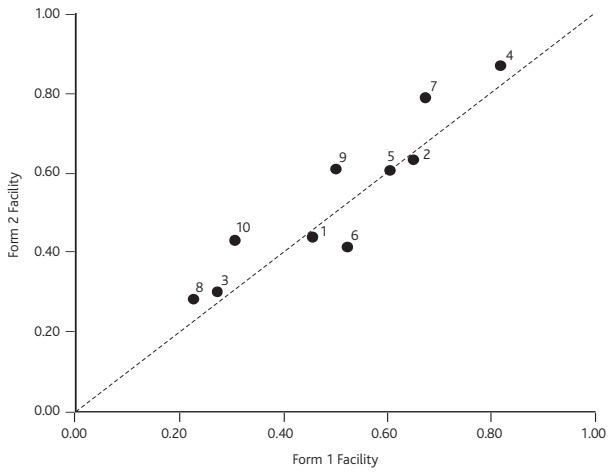


Figure 1: Scatter plot of item facility on each form (Classical Test Theory)

Item Response Theory – Difficulty Analysis

In Item Response Theory, the difficulty of an item is established using a common scale, called a 'latent trait', onto which items can be placed in terms of their difficulty and students can be placed in terms of their ability. The model assumes that the difference between a student's ability measure (on the scale) and an item's difficulty measure (on the same scale) is related to the probability of the student correctly answering the item. The higher the student ability measure is, relative to the item difficulty measure, the greater the probability of the student getting it correct.

The difficulty values for the modified items in each of the parallel forms of the test are shown in Table 3. (In this Item Response Theory analysis the common items were assumed to have identical difficulty so the output for these items is omitted.)

Table 3: Item Difficulty Measures

Item No	*Form 1*		*Form 2*	
	Difficulty	*Standard Error*	*Difficulty*	*Standard Error*
6	0.33	0.29	0.52	0.28
7	-0.37	0.14	-0.75	0.12
8	1.45	0.34	1.24	0.31
9	-1.10	0.22	-1.40	0.26
10	-0.25	0.12	-0.29	0.11
11	0.28	0.20	0.61	0.20
12	-0.73	0.31	-1.32	0.34
13	0.89	0.19	0.78	0.17
14	0.38	0.20	0.12	0.21
15	1.07	0.19	0.55	0.17

Differences in item difficulty are evident; however, the accompanying standard error values indicate that these are not statistically significant. This reinforces the interpretation associated with the item facility analysis findings. A visual representation of the data is shown by the scatter plot in Figure 2. The numerical labels on the data points correspond to each of the *modified* items (1 = Question 6, 2 = Question 7, 3 = Question 8 etc.) and the diagonal dotted line represents item forms of equal difficulty. The scatter plot shows all items were close to the line of equal difficulty in their alternative forms.

Note that the two approaches to measuring difficulty produce similar outcomes, items labelled 3, 8 and 10 emerge as the most difficult, and items labelled 4, 7 and 2 emerge as the easiest, although the rank order of items varies slightly according to which assessment model is selected.

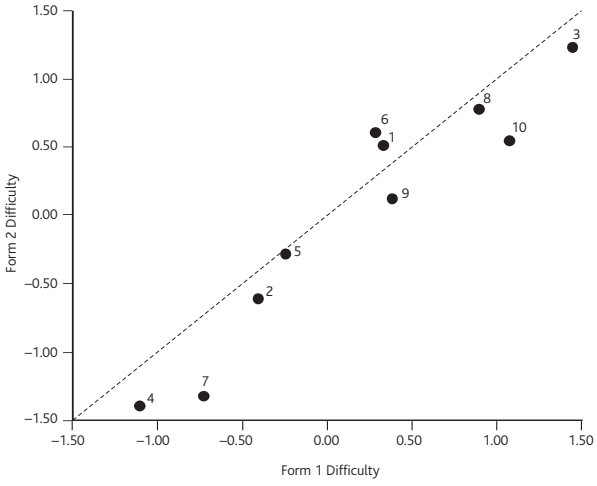


Figure 2: Scatter plot of item difficulty for each form (Item Response Theory)

The impact of ICT competence

The research literature indicated that pupil performance on computer-based tests may be influenced by their competence with ICT (Clariana and Wallace, 2002; Russell *et al*., 2003). Therefore in this study students were asked about their level of competence with ICT. Students were grouped into two subgroups according to their responses: 'ICT High' (n= 43) for those reporting that they were very competent with ICT and 'ICT Low' (n=69) for those reporting that they were less competent with ICT. The difficulty of test items was analysed using these subgroups. Table 4 shows the outcomes of this analysis along with tests for statistical significance.

Table 4: Differences in item difficulty by ICT competence group

ICT Group	*Item no.*	*Facility Form 1 High: n=20 Low: n=35*	*Facility Form 2 High: n=23 Low: n=34*	*Difference*	*t-test statistic*	*Signifi-cance*	*Signifi-cant at 5%?*
ICT High	6	0.46	0.50	-0.04	-0.351	0.726	No
	7	0.68	0.65	0.03	0.542	0.590	No
	8	0.31	0.35	-0.04	-0.336	0.738	No
	9	0.84	0.88	-0.04	-0.532	0.597	No
	10	0.62	0.56	0.06	0.712	0.479	No
	11	0.60	0.35	0.25	2.242	0.029	Yes
	12	0.80	0.82	-0.02	-0.247	0.806	No
	13	0.23	0.28	-0.05	-0.503	0.616	No
	14	0.45	0.63	-0.18	-1.521	0.135	No
	15	0.28	0.44	-0.16	-1.499	0.141	No
ICT Low	6	0.45	0.35	0.10	0.669	0.507	No
	7	0.59	0.61	-0.02	-0.197	0.845	No
	8	0.20	0.22	-0.02	-0.137	0.892	No
	9	0.78	0.85	-0.07	-0.641	0.526	No
	10	0.58	0.67	-0.09	-0.916	0.365	No
	11	0.35	0.50	-0.15	-1.071	0.294	No
	12	0.45	0.74	-0.29	-1.977	0.055	No
	13	0.23	0.28	-0.05	-0.439	0.663	No
	14	0.64	0.57	0.07	0.363	0.720	No
	15	0.36	0.40	-0.04	-0.247	0.807	No

The analysis shows that there were generally no significant differences between the test forms for the two ICT competence groups. One exception was Question 11 which appeared to be significantly different in difficulty for the High ICT group only, with form 2 being significantly more difficult than form 1. This item contained a static artwork in form 1, whereas form 2 was modified to include an animated artwork. The observed difference in difficulty is a curious result that goes against the hypothesis that those with better ICT ability are able to compensate for the ICT demands placed on them. It is possible that the animation was somehow distracting to the high ICT group which meant their responses were not as well thought out.

Discussion and implications

The outcomes indicate that there was little effect on quantitative measures of item difficulty when the item format was changed. Even when the effect of ICT competence on item difficulty was examined, there was very little difference amongst the subgroups. The exception was the anomaly relating to the 'ICT High' group, where the item with animated stimulus appeared to be more difficult than the item with a static stimulus.

It could be argued that the lack of significance observed in the quantitative data means that the item format makes little difference to the difficulty of the item. However, this would be a simplistic implication given the limitations of the context. The sample consisted of 15-year-old students, who may have a very high level of digital literacy compared to the population and this needs to be considered when evaluating the outcomes of this study. There is evidence that poor item design has an impact on the validity of test scores (Huff and Sireci, 2001); therefore, it is important to establish if any of changes to item format would constitute 'poor item design' which would make a difference to the validity of the test.

If large-scale, high-stakes examinations move from paper-based formats to CBA, it is imperative that the effects of item format are well understood to ensure fairness to the students undertaking the assessments. In particular, item design may not have a noticeable effect on the average score in a class of students, but it is possible that individual students may respond very differently to a specific item design.

Limitations

The following indicates areas where generalisations would be more difficult to make, and also suggests areas for further research activity to allow for wider understanding:

- Subject: This study used items assessing the GCSE Science curriculum; it would be useful to understand if other subject areas raised similar findings and issues.
- Item types: This study modified ten aspects of item format; however, the effects may be tied to particular item formats, so a wider study of additional factors could be undertaken.
- Sample: This study was constrained to 15-year-old secondary school students in England, and although a reasonably broad sample of these was achieved, the effects may be different in other student populations.
- The impact of ICT competence: This has not been explicitly explored in this study and there could be more scope for identifying its role in the perceived differences in item difficulty.

Conclusion

The aim of this study was to investigate how aspects of item format in a computer-based assessment affect the difficulty of the test items. All ten aspects of item design that were considered in the study showed no significant difference in measures of item difficulty when administered in parallel forms to the cohort of 112 students. Further investigation could be carried out to look in more detail at the possible confounding effects of ICT competence in this area.

The implications of the study are that the measures of item difficulty appear to be relatively unaffected by the item format presented to the student. However, the computer-based assessments in this study were undertaken by a sample that may have a high level of competence in computer-based applications relative to the general population and this may have affected the findings.

References

Black, T. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design*. London: Sage.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, **16**, 3, 191-205.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, **33**, 5, 593-602.

Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, **32**, 2, 159-181.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research*, **48**, 2, 138-154.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, **12**, 3, 38-47.

Helander, M., Landauer, T. K., & Prabhu, P. V. (1997). *Handbook of human-computer interaction*. 2nd ed. Amsterdam: Elsevier.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, **20**, 3, 16-25.

Kane, M. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. 4th ed. 17-56. Westport, CT: Praeger.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, **20**, 8, 15-21.

McKenna, C. (2001). *Introducing computers into the assessment process: what is the impact upon academic practice*. Proceedings of the Higher Education Close Up Conference. Lancaster, UK: Lancaster University.

Messick, S. (1989). Validity. In: R. L. Linn (Ed.), *Educational Measurement*. 3rd ed. New York: Macmillan.

Popham, W. J. (2000). *Modern educational measurement*. 3rd ed. Boston: Allyn & Bacon.

Ricketts, C., & Wilks, S. J. (2002). Improving student performance through computer-based assessment: Insights from recent research. *Assessment & Evaluation in Higher Education*, **27**, 5, 475-479.

Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education*, **10**, 3, 309-328.

Russell, M., Goldberg, A., & O Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, **10**, 3, 279-293.

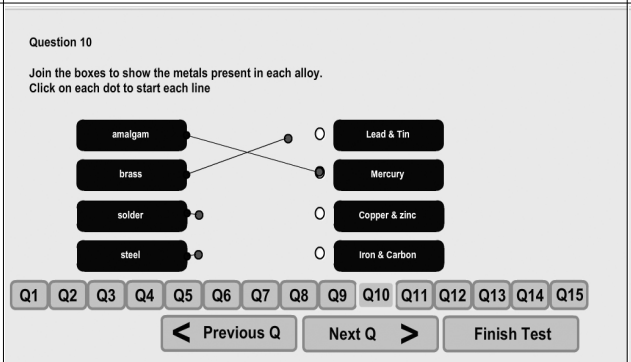
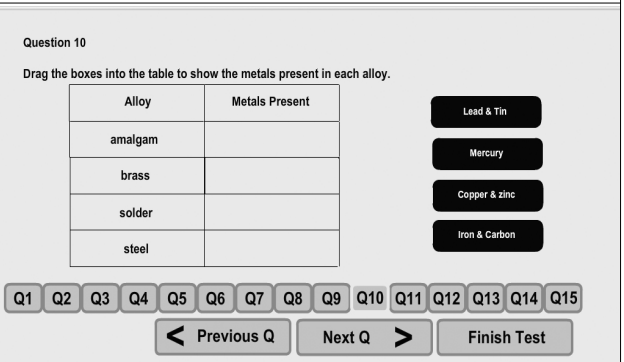
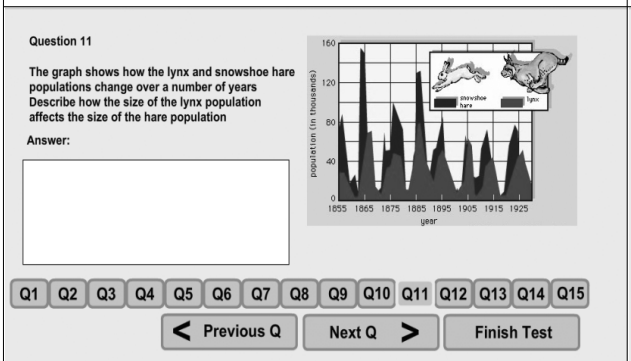
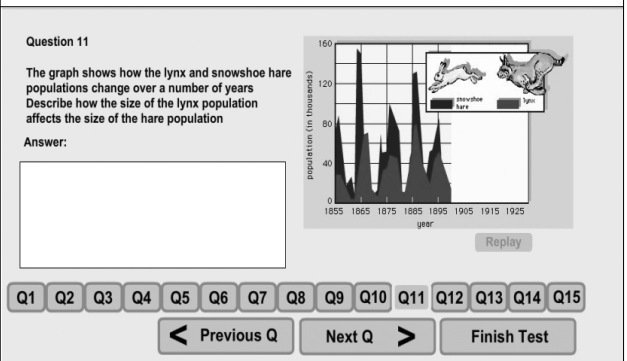
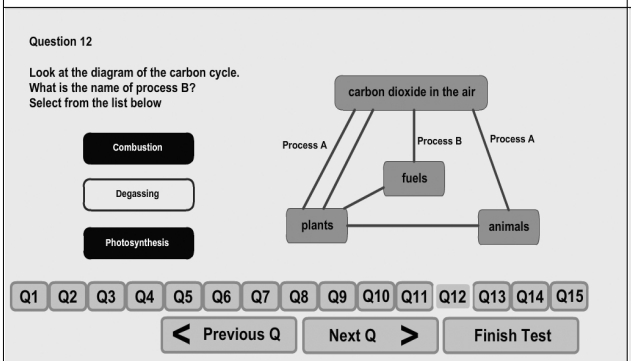
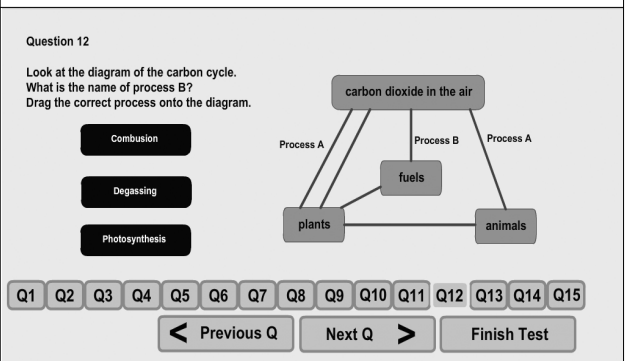
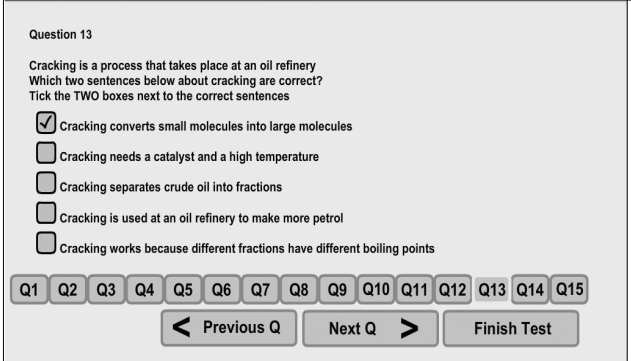
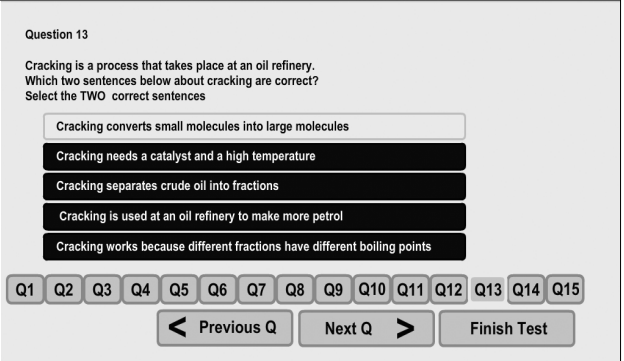
Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In: S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. 329-347. Mahwah, NJ: Lawrence Erlbaum Associates.

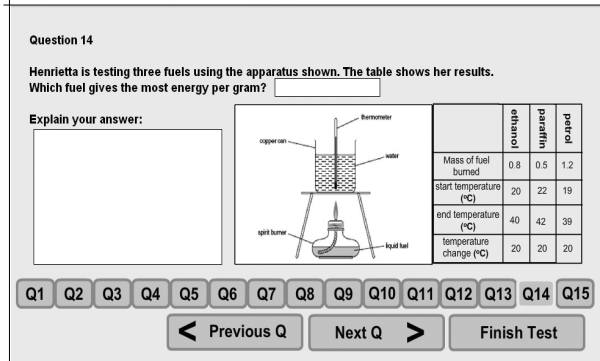
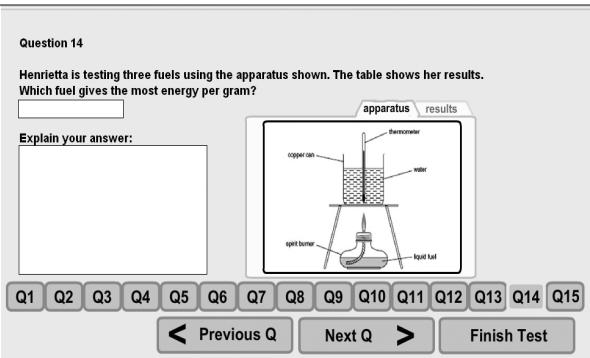
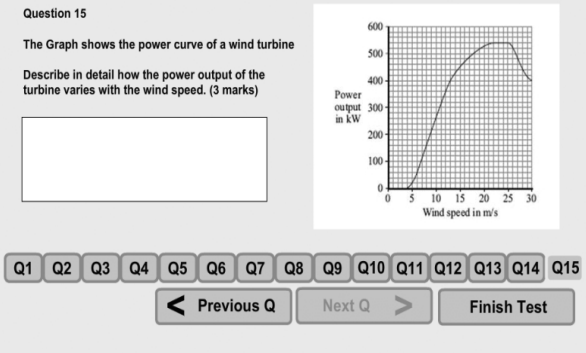
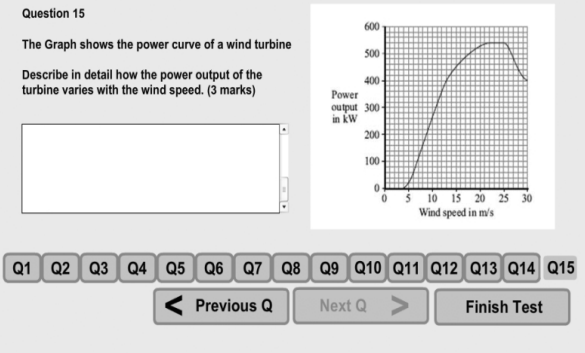
Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide*. Cambridge, MA: MIT Press.

Appendix: Modified test items

Common test items (Q1-Q5) are identical on both forms and therefore omitted.

	Form 1	Form 2
Q6	Supporting colour image provided	No supporting image
	Question 6 Which one of the following is a valid argument for using nuclear power stations? ☐ for maximum efficiency, they have to be sited on the coast ☐ they have high decommisioning costs ☐ they use a renewable energy source ☐ they do not produce gases that pollute the atmosphere  Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test	Question 6 Which one of the following is a valid argument for using nuclear power stations? ☐ for maximum efficiency, they have to be sited on the coast ☐ they have high decommisioning costs ☐ they use a renewable energy source ☐ they do not produce gases that pollute the atmosphere Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test
Q7	Drag and drop categorisation	Tick box categorisation
	Question 7 Human body temperature is controlled in many ways; some of the methods are listed below. Drag each method to the correct column in the table. exercise respiration shivering increase blood flow near skin sweating Ways to Gain Heat | Ways to Lose Heat Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test	Question 7 Human body temperature is controlled in many ways; some of the methods are listed in the table. Tick the box in the correct column for each method. Method | Ways to Gain Heat | Ways to Lose Heat Exercise ☑ ☐ Respiration ☐ ☑ Shivering ☐ ☐ Increase blood flow near skin ☐ ☐ Sweating ☐ ☐ Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test
Q8	Multiple choice, single selection only	Multiple choice, multiple selections enabled
	Question 8 Which of the following is a disease caused by bacteria? ☐ Athlete's foot ☐ Flu ☐ Cholera ☐ Dysentry Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test	Question 8 Which of the following is a disease caused by bacteria? ☑ Athlete's foot ☑ Flu ☑ Cholera ☑ Dysentry Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test
Q9	Drag and drop to fill in the blanks	Drop-down selection to fill in the blanks
	Question 9 When James exercises his breathing rate gets faster. Drag the correct words below to complete the sentence His breathing rate gets faster so that his muscles can recieve _____ more quickly, the muscles also need to remove more _____ carbon dioxide nitrogen protein vitamins oxygen Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test	Question 9 When James exercises his breathing rate gets faster. Drag the correct words below to complete the sentence His breathing rate gets faster so that his muscles can recieve Select... more quickly, the muscles also need to remove more Select... carbon dioxide nitrogen oxygen protein vitamins Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 < Previous Q Next Q > Finish Test

	Form 1	Form 2
Q10	Matching options with lines	Matching objects, drag and drop into table
		
Q11	Static graphic	Animated graphic with replay option
		
Q12	Select option response	Drag answer to target response
		
Q13	Tick-box multiple choice	Selected statement multiple choice
		

	Form 1	Form 2
Q14	All required data presented on one screen	Required data accessed by tabbed panels
		
Q15	Limited text box input	Unlimited text box input with scroll-bar
		

Making the most of our assessment data: Cambridge Assessment's Information Services Platform

Nicholas Raikes Research Division

In *Research Matters*, Issue 12, a report on the evaluation of senior examiners' use of item level data (Shiell and Raikes, 2011) was published. It reported that senior examiners widely used these data and found them helpful for a range of reporting, technical and quality improvement purposes. We also reported that the item level data service was built on a new data and statistical analysis platform in Cambridge Assessment, called the Information Services Platform.

In the present article this new platform will be described in more detail and the thinking behind its introduction explained.

With the Information Services Platform and the innovative strategy it represents, Cambridge Assessment is well placed to take full advantage of the rich data accruing from innovations in assessment technology. It will enable us to better monitor equity and access issues, report on educational trends, assure the quality of our assessments and provide richer and more useful results and information to all our stakeholders.

The promise of better information and the challenges of providing it

One way in which new technologies have the potential to transform large scale educational assessment is by making assessment information more detailed, immediate and accessible than ever before.

For example, consider this list of information-based services that are made possible simply by the relatively straightforward innovation of having traditional scripts marked online and the marks captured at item level:

- Reporting of candidate performance at question or topic level, in addition to reporting at examination and qualification level.
- Extension of item analysis, reliability analysis and statistical screening for malpractice to constructed-response assessments, in addition to the similar analyses long done for machine-read objective tests.
- Near real-time monitoring of marking quality.

While the availability of detailed data is a prerequisite for services such as these, they also depend on data analysis, summarisation and presentation, and it is in these areas that many of the challenges lie.

This article describes these challenges – and Cambridge Assessment's solutions to them – in relation to providing our senior assessors and managers with flexible, dynamic, on-demand statistical information to help them ensure the validity, reliability and timely delivery of our assessments.

A summary of challenges

The pace of technological change is unlikely to abate soon. As we get more experienced at harnessing new data we will develop new uses for it and refine old ones – and new opportunities will continue to be created by innovations in e-assessment. In this context flexible but scalable provision is essential, as is the need to avoid information overload on the part of the users.

Traditionally there have been two main sources of statistical information at Cambridge Assessment:

1. Analysis and reports built directly into our bespoke examination processing system.
2. Custom analysis and reports addressing particular issues and undertaken by statistical experts using statistical software packages on personal computers.

Both of these sources have advantages and disadvantages.

Advantages and disadvantages of built-in analysis and reports

The advantages of building analysis and reports directly into our examination processing system are:

- The analysis and reports are always based on the latest data, and all authorised users have desktop access to them. Unauthorised users (i.e. those without the necessary system permissions) have no access.
- The system is very reliable and dependable, being managed in a data centre in line with formal standards and with change control and disaster recovery procedures.
- The system has sufficient capacity to process large amounts of data quickly.
- Calculated statistics and flags can easily be incorporated into subsequent processes running in the examinations processing system, and are saved.

The disadvantages are:

- Adding new statistics or reports, or making even minor changes to existing ones, is a considerable undertaking, since their impact on the wider system must be fully understood and tested before they can be used.
- All changes and additions must be made by IT developers who may lack statistical understanding or expertise in presenting statistical information clearly, and who must therefore be very closely briefed by the statisticians who do have these capabilities. Also, the IT developers may not have a clear understanding of how the reports and statistics will be used, making it hard for them to understand all the requirements.
- Sophisticated statistical analyses are hard to implement in software and programming languages not designed for this purpose.

Advantages and disadvantages of analyses and reports produced using desktop software

The advantages of having analyses and reports produced by statisticians using desktop statistical software are:

- New analyses and reports can be delivered very rapidly.
- Sophisticated analyses and graphics can be included easily.
- Everything is under the control of the statistician who typically works very closely with the users of the statistics and reports and therefore does not need to explain his or her requirements to a third party developer.

The disadvantages are:

- This method of undertaking statistical analysis and producing reports is not scalable, since automating production is hard or impossible.
- The availability of analyses and reports depends on the availability of the statisticians. Illness at a critical time, for example, could have a significant impact on availability, since specialist statistical expertise is not easily replaced and large numbers of statisticians are not held in reserve to cover periods of absence.
- Data must be extracted from our examination processing system and imported into the statistical software running on the statistician's personal computer. Therefore, the data used in the analysis may not be the latest even when first used, and the resulting statistical information and reports cannot be updated without a further cycle of data extraction and importation.
- Typically, personal computers have less processing power and smaller memories than server computers, resulting in longer processing times.
- Statistical values created on a statistician's computer are hard to read back in to the examination processing system for use in subsequent processes, and may not be saved in an easily re-usable form.

Our solution: the Information Services Platform

Cambridge Assessment's solution to the problem of providing flexible, scalable, dependable and cost-effective statistical analysis and reports is a hybrid system known as the Information Services Platform (the Platform), which combines the resilience and scalability of a server-based architecture with the flexibility and efficiency of having statisticians responsible for creating the statistical content.

The Platform primarily consists of:

- *A data warehouse*
  This contains operational data sourced frequently and automatically from our examination processing system. Statistics calculated on the Platform can also be permanently saved in the data warehouse, where they are available for use in future analyses and reports and also can be read and used by other systems with access to the data warehouse.
- *Statistical analyses and reporting tools*
  These tools are used by statisticians to specify analyses and reports which run on our servers.
- *Automation tools*
  These are used by statisticians to package up analyses and reports for future on-demand use by users of the secure Intranet Portal (see below), or to run them automatically at scheduled times or when specified criteria are met.
- *A secure Intranet Portal*, used by the statisticians for publishing statistical reports and data (content) to authenticated end-users (consumers) across Cambridge Assessment.

Figure 1 is a simplified schematic diagram of the Platform.

The core technology used by the Platform is SAS, which we have long used in Cambridge Assessment as a desktop analysis package. By using SAS technology for the Platform we were able to leverage the advanced SAS programming skills already held by many of our statisticians.

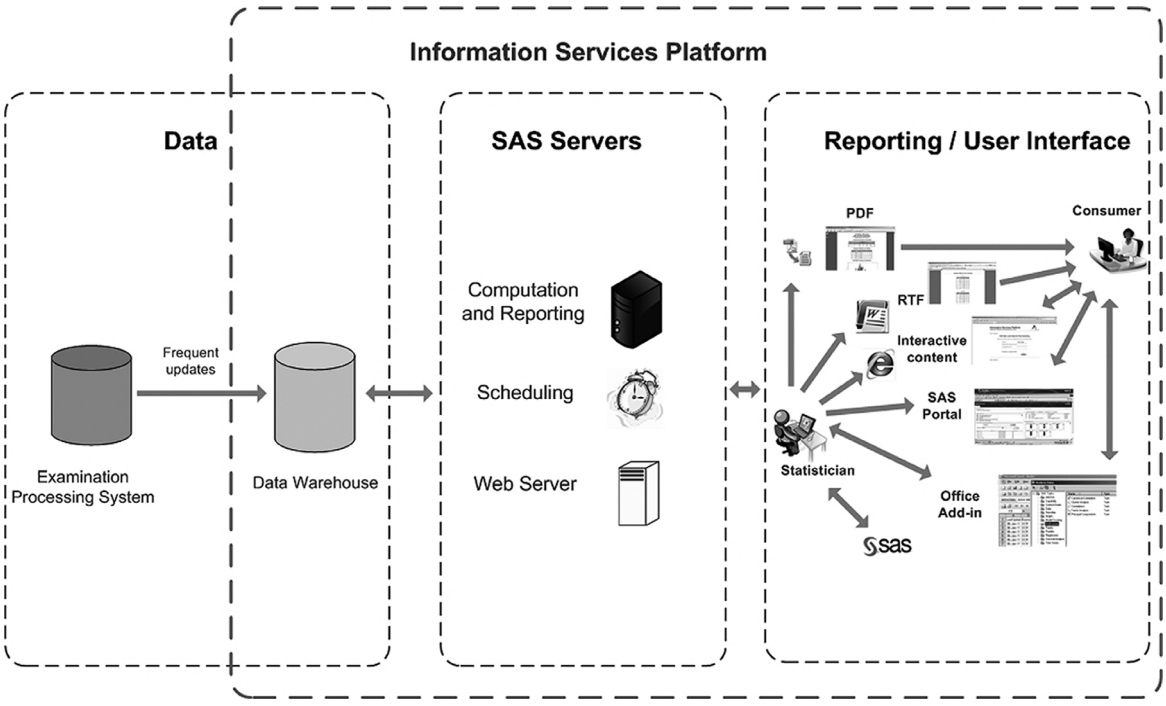


Figure 1: Simplified schematic diagram of the Information Services Platform

Uses of the Information Services Platform

We plan to use the Platform to produce most of the statistical information used by our senior assessors and managers. Current uses include:

- Providing item analyses and reports for all examinations marked on screen;
- Screening marks for signs of candidate or centre malpractice;
- Statistical monitoring of marking accuracy (under development);
- Statistical monitoring of examination comparability (under development);
- Statistical monitoring of examination reliability (under development).

We will also use our data warehouse, combined with data from national databases, to analyse educational trends, equity issues and access issues, and publish the results in authoritative papers and reports, as part of our continuing commitment to providing evidence in support of public policy development and debate.

Conclusion

Evaluation of the Item Level Data service (Shiell and Raikes, 2011) showed that the Information Services Platform enabled our statisticians to provide useful statistical content to senior assessors and managers in a highly reliable, scalable and flexible way. The automated service implemented by our statisticians reliably produced reports for nearly 600 examinations to a tight operational schedule in summer 2010, and has run without major issue for all examination series since then. By using the Platform we are able to combine the flexibility, efficiency and responsiveness of having our statistical experts in charge of creating statistical content, whilst benefiting from the robustness and scalability of a server-based architecture. The Information Services Platform is now a core piece of Cambridge Assessment's infrastructure, central to our vision for taking full advantage of the statistical information made possible by advances in assessment technology.

Reference

Shiell, H. & Raikes, N. (2011). Evaluating Senior Examiners' use of Item Level Data. *Research Matters: A Cambridge Assessment Publication*, **12**, 7–10.

EXAMINATIONS RESEARCH

Statistical Reports

The Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school. These reports, produced using national-level examination data, are available on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports.

Ten new reports have been added to the series since the publication of Issue 12 of *Research Matters*.

- Statistics Report Series No. 29: Predicting A level grades using AS level grades (plus commentary)
- Statistics Report Series No. 30: A level uptake and results, by gender 2002-2010
- Statistics Report Series No. 31: GCSE uptake and results, by gender 2002-2010
- Statistics Report Series No. 32: A level uptake and results, by school type 2002-2010
- Statistics Report Series No. 33: GCSE uptake and results, by school type 2002-2010
- Statistics Report Series No. 34: Provision of GCSE subjects 2010
- Statistics Report Series No. 35: Uptake of GCSE subjects 2010
- Statistics Report Series No. 36: Candidates awarded the new A* grade at A level in 2010
- Statistics Report Series No. 37: Uptake of two-subject combinations of the most popular A levels 2001-2010
- Statistics Report Series No. 38: Uptake of two-subject combinations of the most popular A levels in 2010 by candidate and school characteristics

Research News

Conferences and seminars

International Conference on Thinking (ICOT)

Beth Black presented two papers at the ICOT conference in Belfast in June: i) *Critical Thinking and its impact upon wider academic performance in school*; ii) *An overview of a programme of research to support the assessment of critical thinking*.

Values and Purpose in Citizenship, Social and Economics Education

In June Sanjana Mehta presented a paper entitled: *Why study Economics? Perspectives from 16 to 19 year old students*.

Journal of Vocational Education and Training (JVET)

Martin Johnson and Jackie Greatorex attended the JVET conference in Oxford in July. Martin gave a paper entitled: *Can you dig it? Developing an approach to validly assessing diverse skills in an archaeological context*. Jackie presented a paper on: *Comparing specifications in a diverse qualifications system: instrument development*.

International Computer Assisted Assessment (CAA)

Matt Haigh presented a paper at the CAA conference in Southampton in July entitled: *An investigation into the impact of item format on computer-based assessments*.

British Educational Research Association (BERA)

The BERA Annual Conference was held from 6-8 September 2011 at the Institute of Education, University of London. Colleagues from the Research Division and CIE presented the following papers:

Carmen Vidal Rodeiro: *Do special consideration enhancements skew examination grades?*

Sanjana Mehta, Irenka Suto, Gill Elliott and Nicky Rushton: *Independent research at A level: students' and teachers' experiences*.

Tom Bramley and Vikas Dhawan: *Estimates of reliability at qualification level for GCSE and A level examination*.

Matt Haigh: *An investigation into the impact of screen design on computer-based assessments*.

Martin Johnson, Rebecca Hopkin, Hannah Shiell and John Bell: *Extended essay marking on screen: does marking mode influence marking outcomes and processes?*

Nicky Rushton, Irenka Suto, Gill Elliott and Sanjana Mehta: *Small is beautiful? An exploration of class size at A level*.

Irenka Suto, Gill Elliott, Nicky Rushton and Sanjana Mehta: *Going beyond the syllabus: views from teachers and students of A level Mathematics*.

Victoria Crisp and Rebecca Hopkin: *Modelling question difficulty in an A level Physics examination*.

Victoria Crisp and Stuart Shaw: *How valid is A level Physics? A wide-ranging evaluation of the validity of Physics A level assessments*.

Jackie Greatorex, Nicky Rushton, Sanjana Mehta and Rebecca Hopkin: *Comparing specifications in a diverse qualifications system: instrument development*.

Gill Elliott: *100 years of controversy over standards: making sense of the issues*.

Jackie Greatorx: *Comparing different types of qualifications (e.g. vocational versus academic)*.

Stuart Shaw and Victoria Crisp also presented a poster entitled: *Identifying a set of methods for validating traditional examinations: a difficult task requiring multiple methods*.

European Conference on Educational Research (ECER)

In September Tom Bramley attended the ECER annual conference in Berlin and presented a paper entitled: *Investigating and reporting information about marker reliability in high stakes external school examinations*.

International Association for Educational Assessment (IAEA)

The 37th annual IAEA conference took place in October in Manila, Philippines. The conference theme was 'The assessment and challenge of globalisation'. Nick Raikes presented a paper on *Making the most of our assessment data: Cambridge Assessment's Information Services Platform*.

Association for Educational Assessment – Europe (AEA-Europe)

The AEA-Europe annual conference took place in Queen's University, Belfast in November with the theme of 'Managing Assessment Processes: Policies and Research'.

The following papers were presented:

Carmen Vidal and Sylvia Green: *Linear or modular – does one size fit all? An investigation into the effects of modularisation at GCSE*.

Rebecca Hopkin and Victoria Crisp: *Item difficulty modelling: exploring the usefulness of this technique in a European context*.

Victoria Crisp: *The judgement processes involved in the assessment of project work by teachers*.

Stuart Shaw and Victoria Crisp: *Translating validation research into everyday practice: issues facing an international awarding body*.

Nicky Rushton: *What form of interim feedback most motivates students? A study of teachers' perceptions of the impact of assessment*.

Stuart Shaw and Victoria Crisp also presented a poster entitled: *An argument-based approach to validation: building, evaluating, and presenting the arguments*.

For copies of conference papers please visit our website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers

Publications

In October a Special Issue of *Research Matters* on Comparability was published. This explores some of Cambridge Assessment's recent thinking on Comparability and includes a range of articles on terminology, method, subject difficulty and comparing different types of qualifications. For a copy of the Special Issue please email: researchprogrammes@cambridgeassessment.org.uk or visit the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Research_Matters

The following articles have been published since Issue 12 of *Research Matters*:

Bell, J.F. (2011). The small-study effect in educational trials. *Effective Education*, **3**, 1, 35-48.

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

Crisp, V., Johnson, M. and Novaković, N. (2011). The effects of features of examination questions on the performance of students with dyslexia. *British Educational Research Journal*. Available online at: http://www.tandfonline.com/doi/abs/10.1080/01411926.2011.584964

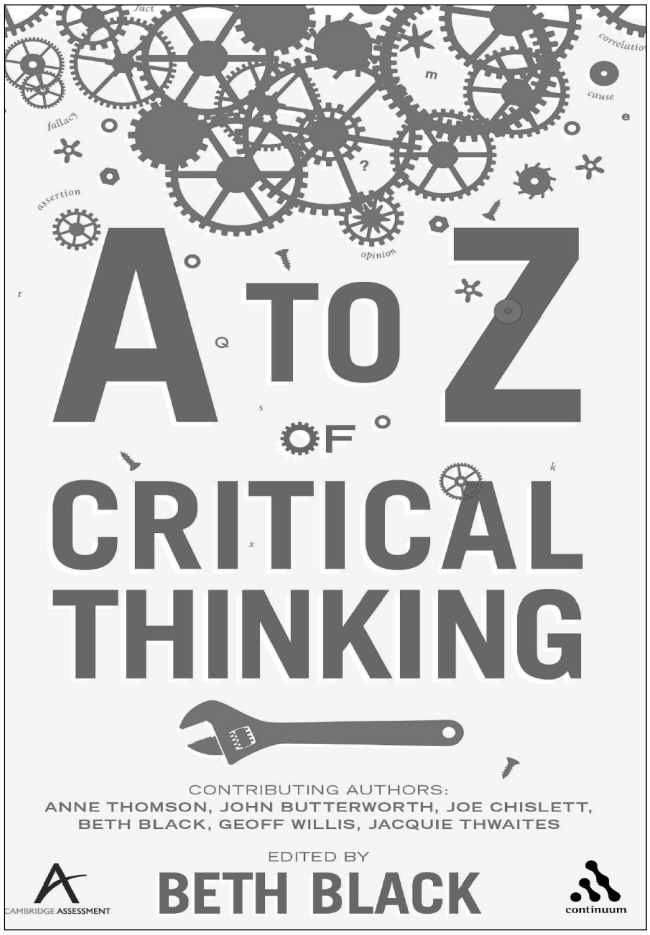
Crisp, V. (2011). Identifying features that affect the difficulty and functioning of science exam questions for all candidates and specifically for those with reading difficulties. *Irish Educational Studies*, **30**, 3, 323-343.

Gill, T. and Bell, J.F. (2011). What factors determine the uptake of A level Physics? *International Journal of Science Education*. Available online at: http://www.tandfonline.com/doi/full/10.1080/09500693.2011.577843

Nádas, R. (2011). Tudod-é? – iskolai e-értékelés (E-assessment in schools). Mindennapi Pszichologia (Everyday Psychology)

CRITICAL THINKING

A to Z of Critical Thinking



A to Z of Critical Thinking, just published in the UK, is a definitive reference tool on Critical Thinking, offering clear explanations and enlightening examples of all the key terms and concepts. This book is the product of a collaboration between the Research Division at Cambridge Assessment and leading experts in this discipline. The book is published by Continuum Publishing, a leading independent academic publisher based in London and New York.

Critical Thinking has become increasingly prominent as an academic discipline taught and examined in schools and universities around the world, as well as a crucial skill for everyday life. A successful critical thinker needs to understand how the different concepts and terms are defined and used. However, Critical Thinking – perhaps more than many disciplines – suffers from problems of definition since much of its terminology is used imprecisely (or just differently) in everyday language.

This definitive A to Z guide provides precise definitions for over 130 terms and concepts used in Critical Thinking. Each entry presents a short definition followed by a more detailed explanation. The aim of this glossary is to provide authoritative clarification and disambiguation of the terms and concepts that are the tools of good thinking.

Cambridge Assessment has been assessing Critical Thinking in a variety of tests and qualifications for over two decades (longer than any other UK awarding body) and, as such, has a special interest in the discipline. This 'A to Z' is the culmination of a longer programme of research, which has included deriving a definition and taxonomy of Critical Thinking (Black *et al*., 2008), exploring how the discipline is taught in schools (Black, 2010), and its impact on other academic disciplines (Black and Gill, 2011).

We hope that students, teachers, academics and anyone wishing to hone their thinking will find some advantage in reading this book and after doing so will be able, for example, to distinguish an assertion from an argument, a flaw from a fallacy, and a correlation from a cause.

A to Z of Critical Thinking is available from bookshops, from online retailers and from the publisher (www.continuumbooks.com). It is available as paperback, hardback and as an ebook.

The book will also be published in the US in February 2012.

ISBNs:
Paperback: 9781441117977
Hardback: 9780826420558
eBook (for individual purchase, Kindle etc.): 9780826436955
eBook (PDF, for institutional purchase): 9781441138422

References

Black, B. (2010). "It's not like teaching other subjects" – the challenges of introducing Critical Thinking AS level in England. *Research Matters: A Cambridge Assessment Publication*, **10**, 2-8.

Black, B. & Gill, T. (2011). Does doing Critical Thinking AS level confer any advantage for candidates in their performance on other levels? *Research Matters: A Cambridge Assessment Publication*, **11**, 20-24.

Black, B., Chislett, J., Thomson, A., Thwaites, G. & Thwaites, J. (2008). A definition and taxonomy for Critical Thinking. *Research Matters: A Cambridge Assessment Publication*, **6**, 30-36.

CONTENTS : Issue 13 January 2012

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031