Research Matters

A selection of articles

CAMBRIDGE ASSESSMENT

Assessment instruments over time

Cookery exams over the last 70 years

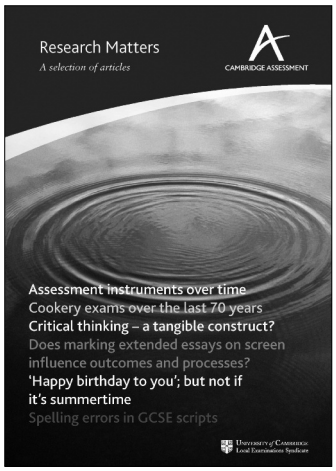
Critical thinking – a tangible construct?

Does marking extended essays on screen influence outcomes and processes?

'Happy birthday to you'; but not if it's summertime

Spelling errors in GCSE scripts

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

A CAMBRIDGE ASSESSMENT PUBLICATION

Introduction

Research Matters is a biannual publication from Cambridge Assessment and this selection of articles highlights some of the areas of research carried out across the organisation. Cambridge Assessment is the largest assessment agency of its kind in Europe and has a key role to play in influencing thinking on assessment. *Research Matters* reports on the detailed and varied research carried out by Cambridge Assessment and shares it with colleagues in the wider assessment community both nationally and internationally. We investigate prominent research issues and report on a range of areas in testing and assessment. From time to time we produce special issues of *Research Matters* which report on particular research that requires a longer, more detailed publication. *Research Matters* is one of a family of publications from Cambridge Assessment, which includes a quarterly publication, produced by Cambridge ESOL, part of the Cambridge Assessment Group. *Research Notes* covers the extensive programme of research, test development and validation work on language assessment carried out within Cambridge ESOL.

In the first article the Evaluation and Psychometrics team marked Cambridge Assessment's 150th anniversary by looking back at question papers over the years. They describe some of the educational and social changes that have affected students over time and illustrate them through changes in question papers from seven subjects. Elliott continues the historical theme in her article on the examination of cookery from 1937 to 2007. Her work provides insights into how the subject has evolved over the years.

In her article on critical thinking Black takes us to a more 'modern' area of study. She engages with a challenging area of assessment in the context of a subject that has proved difficult to define and to measure. She discusses the construct of critical thinking and engages with some of the debates over the last forty years during which interest in this area has increased.

Shiell *et al.*'s article reports on research into modern processes and the influence of marking mode on outcomes and processes. Developments in technology have led to changes in marking processes with examiners marking digitally scanned copies of examination scripts on screen rather than the original paper documents. This research investigates some of the consequences of this shift and is important in answering fundamental questions about onscreen marking in the context of extended writing.

An influential research review is reported in the article on the effect of birthdate on performance. The review from Oates, Sykes, Emery, Bell and Vidal Rodeiro provides robust evidence from around the world that, on average, the youngest children in their year group at school perform at a lower level than their classmates. The review detailed in this article was released to the press in February 2009. It was widely reported in England and received attention in other countries, including China. At the same time it was submitted as evidence to the Rose review of primary education which, as part of its interim report, had recommended that all children should start formal schooling at the age of four (rather than five, as is currently the case). The final article from Elliott and Johnson reports on research into the nature of spelling errors and whether certain spelling errors were particularly common and how they related to spelling conventions, as taught in schools. In their work they discuss the implications of their findings for teaching and literacy policy.

At Cambridge Assessment our research covers a wide range of subjects and in *Research Matters* we report on many technical areas of assessment and measurement. The six articles selected for this issue focus on assessment in more general educational contexts rather than the more technical measurement areas that are also covered in our regular publications.

Sylvia Green *Director of Research*

Assessment instruments over time

Gill Elliott, Milja Curcin, Nat Johnson, Tom Bramley, Jo Ireland, Tim Gill and Beth Black Research Division

First published in Research Matters, Issue 7, January 2009

Introduction

As Cambridge Assessment celebrated its 150th anniversary in 2008 members of the Evaluation & Psychometrics Team looked back at question papers over the years. Details of the question papers and examples of questions were used to illustrate the development of seven different subjects. In each case the following research questions were addressed:

- Has the assessment structure altered over time?
- Have the emphases on different topic areas changed over the years?

The seven subjects studied were:

Mathematics	Physics	Geography	Art
French	Cookery	English Literature	

Background

In the 150 years since Cambridge Assessment/University of Cambridge Local Examination Syndicate has been in existence, there have been a great many educational and social changes affecting students, teachers and assessments. This project sought to describe some of these changes and to illustrate them through changes in question papers. The project was a departure from the usual qualitative and quantitative methods used by the Evaluation Team, and instead took the form of a semi-structured investigation of the development of a number of subjects through the questions presented in the written examination papers.

These studies cannot be used to provide a commentary on *standards* over time, for several reasons:

- First, they do not contain sufficient salient information about the mark schemes, the curriculum and the exact nature of the work produced in response to the questions (scripts). Without *all* of these pieces of information, most of which no longer exist, comparisons about whether a particular era is 'better' simply cannot be made.
- Secondly, examination questions have changed over the years. For example, advances in technology have made it possible to routinely calculate statistics about questions (e.g. facility values) which can provide question writers with important feedback about the performance of that question. Additionally, much development has occurred around question writing and question writer training. Older questions which may seem difficult to 21st century readers may have been difficult for reasons which would nowadays be challenged on the grounds of fairness or validity. Finally, the regulation and oversight of all Awarding Bodies has changed beyond recognition in 150 years. Therefore, simplistically comparing questions from one era with another as evidence of changes in standards over time is flawed.
- Thirdly, the nature of the cohort has altered over the years and examination questions do not show this. So for example, the candidates sitting a School Certificate examination in 1907 might have been only a tiny proportion of the 16-year-old population, whereas the vast majority of 16-year-olds enter for GCSEs in the current context. As a consequence the level of accessibility of the questions differs – modern questions must be worded in such a way that all students being targeted can make some attempt at answering. The target candidature of past questions (particularly those from the earliest years sampled) was undoubtedly very different.

However, studies such as these can be used to illustrate the vast changes that have occurred, and the examples which follow show a small selection of the findings in each subject. These were presented as a poster at the 34th International Association for Educational Assessment (IAEA) Annual Conference which was hosted in Cambridge from 7–12 September 2008 by Cambridge Assessment, as part of the celebrations for its 150th anniversary.

The studies looked at the way in which papers were structured over the years, as shown in these examples from the **Physics** study (Table 1).

Table 1

Year	*Paper*	*Time*	*Rubric*	*Example question*
1927	Physics I	2hrs	Not more than **six** questions are to be attempted.	Explain the phenomenon of dew, and discuss the conditions which favour its formation. How is the dew point determined, and how can the relative humidity of the atmosphere be calculated when the dew point is known?
1957	Physics Ordinary Level Theoretical Paper	$2^1/_2$ hrs	Answer **all** the questions in Part I and **five** questions from Part II including at least **one** question from each of the Sections A, B, C.	{From Part I}: What is the freezing-point of water on the Fahrenheit scale? Express, in °C, a temperature which is 45 degrees below the freezing-point of water on the Fahrenheit scale.
2007	1982/4 Science: Physics extension option A Paper 4 Higher Tier	45mins	Wide range of mark totals per question	This question is about generating electricity. In 2005 the Prime Minister, Tony Blair, called for a 'National Debate' on nuclear power, climate change, and renewable energy sources. (a) Explain what is meant by a **renewable energy source**. [2] (b) More nuclear power stations could be built. (i) Suggest **two** arguments for building more nuclear power stations. [2] (ii) Suggest **two** reasons **against** building more nuclear power stations. [2]

The two key themes which have been identified across many of the subjects include the increase in the number of questions relating to real-world contexts, and the greater amount of choice available to candidates, both in terms of the different options within assessments and the methods by which they may display their skills.

Increasing use of real-world contexts can be illustrated from the study into **Mathematics**, where it was interesting to note that as early as 1957 one of the regulations sections stated that some of the questions might be set on the application of certain arithmetical processes to problems of everyday life in the home and the community. This appears to be one of the early explicit statements indicating a trend that became prevalent in testing all topic areas of mathematics in the GCSE Mathematics papers, although it was present even in the 19th century papers to some extent, especially in the area of Arithmetic.

An example of a question from the 1997 GCSE Mathematics assessment:

Mrs McKenzie bought a large box of bags of crisps for her family. She told the children that the box should last 3 weeks if they ate 12 bags per week between them.

(i) How many weeks should the box last if the children eat 9 bags per week between them?

If the children eat n bags per week between them, the box will last W weeks.

(ii) Write down a formula which connects W and n.

The studies investigated how topic areas within subjects have altered over the years. In this example from the **Geography** study (Figure 1), physical geography, human geography and geographical skills have featured since early days, but economic and environmental geography are more recent elements of the assessment.

In some instances practical considerations have affected the practice of assessment.

Photograph by Peter Askern. Cambridge Assessment Archives Ref: M/P 5/8

Figure 2: Artwork in 1 Hills Road for marking

For example, **Artwork** (Figure 2) used to be necessarily restricted by weight and size, because the work was sent to Cambridge and displayed in the Craft Hall at 1 Hills Road for marking.

"Pieces of pottery must not exceed 12 ins. in any dimension, nor exceed 7 lbs. in weight. Pieces of sculpture or carving must not exceed three feet in any dimension nor exceed 20 lbs. in weight."
1977 and 1987 Art specification

Now that schools themselves display candidates' work and examiners make visits to the schools, students' artwork is not limited in this way.

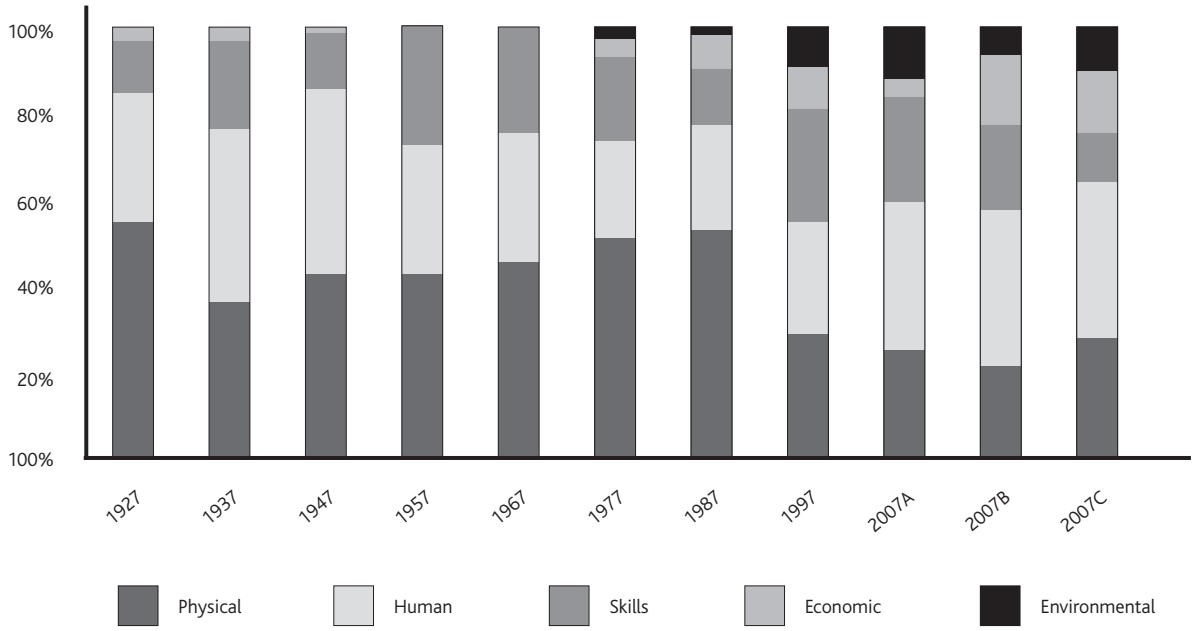


Figure 1: Geography: Summary of topic areas over time

Table 2: Skills tested over time in English Literature

Skill	*1877*	*1887*	*1897*	*1907*	*1917*	*1927*	*1937*	*1947*	*1957*	*1967*	*1977*	*1987*	*1997*	*2007*
grammatical analysis	✔		✔											
etymology	✔	✔	✔											
textual analysis	✔	✔	✔	✔										
scan (divide into metrics), knowledge of poetic/linguistic form (pentameter)		✔	✔	✔		✔	✔							
knowledge of author's life (external to text)			✔	✔		✔								
produce quotations verbatim	✔	✔	✔	✔	✔	✔	✔							
knowledge of literary, dramatic or poetic terms, concepts and mechanisms	✔	✔	✔	✔	✔	✔	✔	✔	✔					
translate text into contemporary prose retaining exact meaning	✔	✔				✔	✔		✔	✔				
comparison of text with factual information/external point of reference			✔	✔	✔	✔	✔				✔			
explain meaning of (extended) text (expound)	✔		✔	✔	✔	✔	✔		✔	✔		✔		
exact context of quote/excerpt	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔		
overall evaluation of play/text/poem		✔	✔				✔	✔		✔	✔	✔		
give an account of a scene/sequence of events/story strand/poem	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	
character analysis/development including comparison of characters	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
thematic analysis/overall theme					✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
qualities of the writing of the poem/text						✔	✔	✔	✔	✔	✔		✔	✔
views or attitudes of the author as detected from the text.						✔	✔	✔	✔		✔	✔	✔	
appreciation of text/describing impact upon self/reader response						✔	✔	✔	✔	✔	✔	✔	✔	
significance (importance) of a feature or scene of text							✔	✔				✔	✔	
understanding of staging of play/dramatic impact							✔		✔				✔	
reading comprehension of text (within 'extract based questions')											✔	✔	✔	✔
relating to characters											✔	✔		
imaginative writing – role playing a character in the text											✔	✔	✔	✔
unseen poem											✔	✔		

Finally, the studies have enabled analysis of the skills required by candidates at different points in time in specific subjects. In **English Literature** every question paper between 1877 and 1937 inclusive (Table 2) required candidates to quote verbatim from memory fairly substantial sections of the prescribed text. Earlier question papers used to require candidates to know the precise meaning, usage and etymology of words in the texts, and on occasion, questions would require candidates to quote a line in which a particular word appeared. Later question papers gave more emphasis to discussing overall meaning or themes of a text and describing or analysing the candidate's own response to a passage or character. A particularly common feature of later papers asked candidates to imaginatively play the role of a character in the text.

Summary

The research proved a very interesting means of investigating the development of individual subjects. Naturally the method used – sampling question papers from every tenth year- has some limitations. It is, for example, possible that short-lived topics or question paper structures have escaped our attention altogether. Also the researchers are unable to state for certain exactly when a particular change occurred – the research shows merely the first sampled year when such changes were seen.

However, many interesting details have emerged from every subject studied and two themes were repeated across many of the subjects. These were an increasing emphasis upon real-world contexts for questions in more recent years, and an increasing choice of topic areas and question/component options available to candidates.

For full reports in each of the seven subjects, please contact Gill Elliott, Assessment Research and Development Division, Cambridge Assessment, 1 Regent St, Cambridge, CB2 1GG. Email: elliott.g@cambridgeassessment.org.uk.

Cookery examined – 1937–2007: Evidence from examination questions of the development of a subject over time

Gill Elliott Research Division

First published in Research Matters, Issue 6, June 2008

Introduction

The teaching of cookery skills in UK schools has become the subject of much debate in recent years. Like its counterpart, needlework, the subject has a history of social change and gender bias. In the early twentieth century, when school examinations began to become widespread, both subjects were highly used in a domestic context. In other words, they were life skills, for at least some part of the population. Initially, undoubtedly, both cookery and needlework were subjects undertaken by girls, in the same way as woodwork and metalwork were 'for' boys. In the 1970s and early 1980s there was more integration of boys to the subjects. However, as school subjects, they became increasingly a minority option by both sexes, until they almost disappeared altogether in the 1980s.

As we approach the end of the first decade of the twenty first century, needlework remains a minority option at GCSE, mostly taken by girls (across all awarding bodies in 2006, 45,950 girls took the textiles option of Design & Technology GCSE as opposed to 1,515 boys) and is no longer necessary to any individual as a 'life skill' – nobody suggests that the 21st century family should return to making a substantial number of their own clothes, as was commonly the case into the 1950s at least.

Cookery, however, has been the subject of a recent backlash, with increasing calls for a return to 'traditional' home cooking, with its allied skills of budgeting and planning. The concern has been driven by a number of issues and campaigns – obesity, crises in the NHS, animal welfare debates, environmental concerns surrounding packaging and wasteful food management and the key issue of the long term effects on human health of a diet based largely upon heavily processed foods. As a result, concern is growing that the skills necessary to prepare nutritious well-balanced meals from fresh ingredients have been lost to large parts of the population in a domestic context, and are at critical point within schools.

The purpose of this article is to take a step back from the increasingly heated debates surrounding the state of the UK's diet, and use evidence from the questions set at GCSE over the years in one examination board to look at how the subject has evolved within schools over the years.

The terminology used to describe the subject has changed significantly during the years. As far as possible, in this report, the terms used are those used commonly in schools to describe the subject. Therefore, 'cookery' is used to describe the school subject taught from the 1940s until the 1980s. From the 1990s onward, 'food' has been used as a common generic term to describe the subject – e.g. job advertisements can ask for a 'teacher of food', and is used in this context within this report. Examinations are referred to by their title.

There has been a great deal of debate upon this subject, records of which are mostly contained within newspaper articles. Academic research into the subject is less readily available, although it does exist. Dena Attar's book on gender effects of the subject (Attar 1990) and the Moray House College of Education study into how attainment should be assessed within home economics (Cumming *et al*., 1985) are prime examples. However, little of this literature considers the important question of how cookery examinations have changed over the past few decades. Bearing this in mind the focus of this article is how cookery examinations have evolved over the past seventy years.

Historical background

The first record that the Cambridge Assessment archive[1] has for cookery is in 1929, when it appears at School Certificate. In 1927 only needlework and hygiene are listed. Undoubtedly in this era it was a subject for girls only. Indeed at this time it was only a part of a subject – the School Certificate subject of housecraft allowed students to choose two subjects from four – needlework, laundrywork, cookery and housewifery. Cookery became a subject in its own right with the advent of the O level in 1951.

During the 1950s, 60s and 70s the examination title remained as 'cookery', in this board at least, although the term 'domestic science' was being used in schools and introduced an element of technicality with the use of the word 'science'. Was this an attempt to remove the 'life skill' element and create an academically oriented face to the subject? In the late 1970s the school subject was renamed again, 'home economics', and arguably changed focus from 'teaching working class children the basics of service' to making 'basic and unattractive things with the cheapest possible ingredients' (Purvis, 2007). The title 'home economics', again, uses a term (economics) suggestive of academic rigour.

During the latter part of the twentieth century home economics began to find a place in the craft, design and technology (CDT) suite of subjects, which encompassed electronics, engineering and graphics, as well as wood and metal work and needlework (then called textiles and dress). In the mid 1980s there was consternation when the draft criteria for the subject were rejected by the Secretary of State for Education, because of disagreement about how the new course should be taught and what it should contain (Christian-Carter, 1985). In 1990, according to Geoffrey Thompson of the National Association for Teachers of Home Economics and Technology (reported in Purvis, ibid), the subject of Home Economics was close to being abolished as a method of cutting educational costs. The solution –

1 Cambridge Assessment currently comprises three awarding bodies: OCR, CIE and ESOL. In the past examinations were presented under other names – MEG (Midland Examining Group) and UCLES (University of Cambridge Local Examinations Syndicate). Additionally, other awarding bodies have merged with UCLES, including UODLE, OCSEB and EMREB (see Raban, 2008 for more details). The examination papers discussed in this study are taken from OCR, MEG and UCLES.

hard fought by supporters of the subject – was to ensure that it was contained within the newly created D&T suite, because that was compulsory on the curriculum. Thus 'food technology' became one of the four areas (food, textiles, resistant materials [woodwork & metalwork] and systems [electronics and more]) within the D&T curriculum when the National Curriculum was set up in 1992, and it continues in this form to the current day. However, an alternative home economics qualification has also remained available via several awarding bodies throughout these same years.

Much of the catalyst for the current focus on food in schools came from a TV series – 'Jamie's School Dinners', which aired in 2005 (Channel Four Television, 2005). In the programme, chef Jamie Oliver highlighted the poor state of school dinners in the UK and attempted to change the eating habits of schoolchildren in specific schools. The programme was to a greater extent responsible for a widespread change to the provision of food in schools, including the reduction of 'junk' food availability and an increase in fresh healthy produce (BBC news online, September 2005). The impact of the series was not only a change to school meals but a more widespread concern, about the choices that students and their parents were making about food. It was felt that not only were students being fed over-processed food at school, they were not being educated – either at school or at home – about healthy diets or about fresh ingredients, and what to do with them.

A number of other studies have highlighted a growing crisis in cookery skills/food choices of young people. A study carried out in Scotland emphasised the decline in skills (Horne & Kerr, 2003, reported in McBeth, 2005). In March 2006 Ofsted produced a report on the effectiveness of provision in secondary schools for food technology. It was based upon a survey of thirty secondary schools which taught food technology. The report acknowledged that there had been many concerns raised with inspectors and government officials about the teaching of food in the curriculum in the years preceding the study and, specifically, that the D&T based food technology course emphasised knowledge of food processing and manufacturing at the expense of traditional family cookery. Both the Design and Technology Association (DATA, 2005) and the Children's Food Campaign (Children's Food Campaign, 2006) have advocated the maintenance of food within the curriculum as a matter of priority. 'Every Child Matters: Change for Children' (HM Government, 2005), cited the rights of children to lead a healthier life and to develop skills for living. As a result, provision in schools will change from 2008. In the Design and Technology Association briefing paper for members (DATA, 2007) which summarises the changes, the introduction states that:

> *For those of you that have been in teaching for a number of years, it has been a long struggle for the value of food teaching in a broad education to be recognised and to become highest priority in our schools.*
>
> *This year sees a number of essential education programmes uniting to change the future of children's health and well-being to reinforce the changes that have taken place through 'Jamie's School Dinners'.*

A new KS3 programme of study is described by QCA (QCA, 2007), with the goals of teaching, 'a broad range of practical skills, techniques, equipment and standard recipes' to learn to 'carry out a broad range of practical cooking tasks safely and hygienically', to study healthy eating models and balanced diet, and to learn about 'the characteristics of a broad range of ingredients, including their nutritional, functional and sensory properties'.

At KS3, in the revised National Curriculum, food was not compulsory, although resistant materials, systems and control were. This raised concerns from the Design and Technology Association, not least because of the potential for gender inequality. In January 2008 Ed Balls, the Secretary of State for Children, Schools and Families, announced that from 2011 all schools must offer a food technology curriculum at KS3, with the allied training of 800 new cookery teachers (DCSF, 2008).

'Licence to Cook' is a compulsory cooking entitlement for each student. This will be brought into schools from September 2008, although those schools offering food at KS3 will automatically meet the criteria imposed, which match the KS3 programme of study goals. 'Licence to Cook' will be run by a consortium of three associations: the British Nutrition Foundation, Design and Technology Association and Specialist Schools and Academies Trust.

At KS4 changes are also planned. Awarding Bodies will be required to use the same core competencies to underpin specifications as used at KS3 and 'Licence to Cook'. This is likely to mean less focus on industrial processes at GCSE.

To what extent can Cambridge Assessment provide evidence with which to inform this debate? Table 1a shows the nature and structure of qualifications offered at age 16 by OCR and its predecessors every tenth year from 1937 to 1987, during the period when a single qualification existed. Table 1b continues the table from 1987 to 2007 with the home economics qualification and Table 1c with the D&T food technology qualification. Tables 2a–c provide example questions from the examinations, arranged in the same way. The tables show the information that could be obtained from the question papers – the nature of questions and the structure of the paper. Information about marks allowed, weightings of papers and the marking of individual questions is not contained within the tables, because it was unobtainable for most examinations prior to the 1970s and 1980s.

Discussion

Evolution of the examination

A number of similarities – and differences – between the examinations become apparent when the tables are studied. There is a clear and distinct evolution of the subject, when we look at the structure of the examination.

In the 'early' years – the 1930s and 1940s – the qualification was only available as an optional part of the wider subject of 'housecraft', which included laundry-work, dressmaking and general housewifery, as well as cookery. Each of the options was presented as a separate section of the written paper, and had a separate practical examination, and therefore candidates taking this option took a single written examination in cookery, and a practical component. Questions on the cookery section of the written paper covered areas including menu planning, choosing particular ingredients, the advantages of different methods of cooking, describing common cookery terms, questions related to practical cookery and nutrition. The practical session involved a planning session, followed by a practical cookery examination, in which candidates were required to prepare a number of dishes that might commonly be served in the home environment. There was no evidence about whether the costs of ingredients for examinations (or for lessons generally) were met by the candidates or the school, or were in some way centrally funded.

Table 1a: The nature and structure of examinations offered by OCR and its predecessors (MEG/ UCLES) every tenth year from 1937 to 1987

	Structure of written paper	*Practical paper/coursework*
1937 Half a School Certificate subject; **Subject title:** Housecraft **Paper details:** 1 section of written paper 1 practical paper	45 minutes for the cookery section. One written paper section (presented in combination with Laundrywork, Housewifery & Needlework). Two questions to be answered from a choice of three. Questions multi-part.	Two and a quarter hours. One task allotted to the candidate. No preparation time indicated, nor any indication of candidate having advance notice of dishes to be cooked. Tasks included the preparation of three to five dishes.
1947 Half a School Certificate subject; **Subject title:** Housecraft **Paper details:** 1 section of written paper 1 practical paper	One hour for the cookery section. One written paper section (presented in combination with Household Management & Needlework). Between two and four questions to be answered from a choice of five. Questions multi-part.	One hour planning session. Candidates were given the test allocated to them, and planned what they wished to cook. They had to draw up a plan of work and a list of ingredients. All work was handed in at the end of the planning session and was returned to them at the examination. Candidates had to keep to their written plan of work during the examination, which lasted two hours. Tasks mostly contained three main dishes, plus a small accompaniment – i.e. a drink, or a sauce. Two hours were allowed to complete the task.
1957 O Level **Subject title:** Cookery **Paper details:** 1 written paper 1 practical paper	Single two hour theory paper. Five questions to be answered. Questions were divided into two sections. Section A (where candidates were advised to spend 25% of time) had a choice of 2 longish answers; candidates had to answer one. Section A questions were often (but not always) synoptic in nature, containing a requirement to describe the scientific/ nutritional background to a given situation and then to plan meals accordingly. e.g. *State in detail the importance of protein in the maintenance of good health. What important points should be borne in mind when choosing protein foods for:* *(a) elderly people;* *(b) vegetarians?* *Plan meals for one day for an elderly couple living on a pension and underline the foods which are good sources of protein.* Section B had 6 multi-part question choices of which candidates had to answer four.	One hour **and ten minute** planning session. A choice of two tests was given to each candidate, and they had ten minutes in which to choose which one to take. Candidates then spent one hour preparing a plan of work and a shopping list. Everything was handed in at the end of the planning session and was returned to them at the examination. Candidates had to stick to their written plan of work and might not bring any additional notes (except recipe book). Tasks contained three or four main dishes – sometimes more smaller dishes. Two and a quarter hours allowed for cooking.
1967 O level **Subject title:** Cookery **Paper details:** 1 written paper 1 practical paper		One hour **and a quarter** planning session. Otherwise as 1957 above
1977 O level **Subject title:** Cookery **Paper details:** 1 written paper 1 practical paper		One hour **and a half** planning session. Otherwise as 1957 above
1987 Joint O level/CSE **Subject title:** Home Economics: Food & Nutrition. **Paper details:** 1 written paper 3 practical assignments	2 hour theory paper presented as two sections. Books containing *recipes only* were permitted. Section A consisted of ten compulsory short answer/multiple choice questions. Section B presented two structured, two data-response and two free response questions. Three questions had to be attempted, one from each part.	Three practical assignments. First assignment: a food based problem with one factor, set by teacher. Second assignment: a piece of investigation, set by teacher. Third assignment: a complex problem with two main factors, chosen by the candidate from three assignments set by the Board. Each of these carried out within 2 hours and 15 minutes, spread over 2 weeks, (1 hour planning, 1 hour executing (usually a week later) and 15 minutes evaluating).

Table 1b: The nature and structure of examinations offered by OCR in Home Economics: Food & Nutrition from 1997 to 2007

	Structure of written paper	*Practical paper/coursework*
1997 GCSE **Subject title:** Home Economics: Food. **Paper details:** 1 written paper 3 practical assignments 2 hour theory paper presented as two separate sections.	Section A consisted of ten compulsory short answer/multiple choice questions. Section B presented two structured, two data-response and two free response questions. Three questions had to be attempted, one from each part.	Three practical assignments. First assignment: a food based problem with one factor. Second assignment: a piece of investigation. Third assignment: a complex problem with two main factors, chosen by the candidate from three assignments set by the Board. Each of these carried out within 2 hours and 15 minutes, spread over 2 weeks, (1 hour planning, 1 hour executing (usually a week later) and 15 minutes evaluating).
2007 GCSE **Subject title:** Home Economics: Food. **Paper details:** 2 written papers comprising 1 Foundation and 1 Higher tier. 3 practical assignments	One theory paper to be taken by each candidate. All questions on both papers are compulsory. Both papers contain short answer, structured, data response and free response questions.	Three tasks: One investigative task – 12–14 hours. Two resource tasks 'short focused tasks with the emphasis on the implementation of practical skills'. Each task should take 2–3 hours, and it is expected that 'a number' are conducted throughout the course, but only two be submitted for the assessment.

Table 1c: The nature and structure of examinations offered by OCR in D&T Food Technology from 1997 to 2007

	Structure of written paper	*Practical paper/coursework*
1997 GCSE **Subject title:** Design & Technology Syllabus A: Food Technologies **Paper details** 1 written paper 2 coursework tasks Plus 3 other syllabuses available within D&T suite.	Two compulsory theory papers. Part A: Core (basic tier 45 minutes each, standard tier 1 hour each & higher tier 75 minutes each) contained compulsory structured questions on the core content. Part B: Compulsory structured questions on the optional content.	Two coursework tasks, each taking around 20–30 hours to produce. **One piece of work must demonstrate the use of construction materials i.e. wood, metal, plastic, clay and components.** **The other piece of work must demonstrate the use of one other material, chosen from graphic media, food or textiles.** No specimen/exemplar assignments could be found. Evidence of achievement was taken from design folders and the artefact.
2007 GCSE **Subject title:** D&T: Food Technology **Paper details:** 4 written papers, comprising 2 Foundation and two higher tier. Coursework.	Two theory papers to be taken by each candidate. Foundation tier candidates had 1 hour for each paper, higher tier candidates had 1 hour 15 minutes. Papers 1/2 contained a product analysis question on any theme. Papers 3/4 contained a product analysis on the published theme for the year, which for 2007 was 'frozen food'. All papers contained short answer/data response type questions.	The coursework consisted of the creation of a three dimensional product, plus a portfolio of supporting material. The portfolio must include the identification of a consumer need, the formulation of a design brief to meet that need, research into and around the brief, the generation of ideas and development of a product, plus evidence of the evaluation and testing of the finished product. The specification recommends a maximum of 40 hours work to be spent on the coursework.

Table 2a: Example questions 1937–1987

Year	*Example questions from the written paper(s)*	*Example assignments from the practical/coursework*
1937	*Compare and contrast boiling and steaming as methods of cooking vegetables. Which do you consider the better method? Give reasons for your choice.*	*Make a pulse soup; show two ways of cooking batter, one as a savoury and one as a sweet; make some scones.*
1947	*Enumerate the advantages of steaming as a method of cooking. By means of labelled diagrams, show **three** methods of steaming. Give **two** examples of foods which may suitably be steamed in each of the ways illustrated.*	*Show your skill in cookery by using batter, short crust pastry, and the creaming method to prepare three dishes. A suitable sauce should be served with one of the dishes.*
1957	*What do you understand by the term 'edible offal'? Name **four** examples and state **one** method of cooking suitable for each. Give clear directions for the preparation, cooking, and serving of a dish containing liver or kidney suitable for a quickly prepared midday meal. What would you look for in choosing the liver or kidney?*	*Prepare and serve a special tea for the headmistress and two visitors to your school. It should consist of dainty sandwiches (two savoury fillings), scones, tea and also a Victoria sandwich and a few small cakes, both made from one basic mixture.*
1967	*What is meant by 'fermentation'? Give the ingredients for and method of making a loaf of bread, using $^1/_2$ lb flour. What are the changes which take place while the loaf is baking?*	*a) Prepare a two-course family dinner for three people. The main course should show an interesting method of cooking inexpensive meat and the preparation, cooking and serving of a fresh green vegetable.* *b) Make some interesting biscuits (using not more than 4oz. flour) and serve them on a tray with coffee.*
1977	*a) What advantages are there in making and baking in large quantities?* *b) Give the basic recipe for making:* *bi. shortcrust pastry using 400g or 500g (1lb) flour;* *bii. a creamed mixture using 200g or 250g ($^1/_2$ lb) self-raising flour.* *c) describe briefly how each mixture could be used to make **three different dishes**.*	*a) Prepare, cook and serve a two course mid-day meal for a family of three, one of whom is on a light diet after an illness.* *b) Use some seasonal fruit to make a small quantity of jam or make some lemon curd.*
1987	(Section B – free response): *Your headteacher is concerned about the amount of so-called 'junk food' eaten by young people today. Evaluate the part 'junk food' plays in their diet and comment on the need for thinking carefully about food and health.*	Third assignment: *The use of convenience food in our diet is increasing.* *a) suggest dishes which show the sensible use of convenience food.* *b) As part of your planning explain how the dishes you have chosen take this point into consideration.* *c) Draw a chart to show how you would compare a home-made dish with the same convenience food dish.* *d) Make a selection from your choice in (a).* *e) Evaluate the outcome.*

Table 2b: Example Home Economics questions 1997–2007

Year	*Example questions from the written paper(s)*	*Example assignments from the practical/coursework*
1997 Home Economics: Food	(free response): *Technology has brought about considerable changes for the consumer. Using the following headings, together with your own ideas, explain how the consumer has gained from these changes.* *a) In the range of food available.* *b) At the supermarket checkout.*	(from specimen assignments) *Children need a balanced diet in order to grow up in good health. Prepare a selection of dishes suitable for children under 5 years.* *a) What are the essential requirements of a child's diet?* *b) Write about the dietary needs of children including any special information.* *c) Suggest some suitable dishes and make a selection which you could prepare giving your reasons for choice.* *d) Plan a course of action.* *e) Carry out your plan.* *f) Evaluate the whole assignment.*
2007 Home Economics: Food & Nutrition	(common question to both tiers): *Food eaten at school is an important part of a teenager's diet. Describe the nutritional requirements of teenagers. Explain how schools can help meet these requirements in the provision of food and drink.*	Resource task *Low fat spreads are often used for spreading onto toast or onto bread when making a sandwich.* *a) Plan a test to look at the spreadability of low fat spreads compared to margarine or butter.* *b) Carry out the test.* *c) Evaluate which is the most suitable and why.*

Table 2c: Example D&T Food technology questions 1997–2007

Year	*Example questions from the written paper(s)*	*Example assignments from the practical/coursework*
1997 Design & Technology Syllabus A	(Part B, basic tier): *Sauces and toppings are often used to make fish dishes attractive to young children. Give **three** reasons why sauces and toppings make fish dishes more appealing. Name a suitable sauce for a child's fish dish. List the ingredients and explain the process needed to make it.*	Coursework requirements. *One piece of work must demonstrate the use of construction materials i.e. wood, metal, plastic, clay and components.* *The other piece of work must demonstrate the use of one other material, chosen from graphic media, food or textiles.*
2007 D&T Food Technology	Paper 2 – Higher tier. *A food manufacturer produces a savoury flan in a test kitchen. The basic ingredients are listed below [list of pastry ingredients & list of filling ingredients]. Describe one different performance characteristic (function) for each of the following ingredients when used in the savoury flan. (i) plain flour, (ii) fat, (iii) egg.* *Further research by the food manufacturer has identified a gap in the market for a new type of savoury flan. The new savoury flan should meet the following specification: reflects a culture or a country; combines a variety of different textures in the filling; is attractive in appearance. Complete the chart to describe how the basic ingredients could be adapted to meet the specification.* *Identify one pre-manufactured component which could be used in the new product. Give **two** benefits to a manufacturer of using pre-manufactured components. Give **one** limitation to a manufacturer of using pre-manufactured components.*	The coursework consisted of the creation of a three dimensional product, plus a portfolio of supporting material. The portfolio must include the identification of a consumer need, the formulation of a design brief to meet that need, research into and around the brief, the generation of ideas and development of a product, plus evidence of the evaluation and testing of the finished product.

From the 1950s to the 1970s (the O level era) the subject formed an entire qualification. The practical examination continued in the same format as in previous years (a planning session, followed by a practical cookery examination, in which candidates were required to prepare a number of dishes that might commonly be served in the home environment), albeit with the planning session being given greater time allowance with every successive decade. The theory paper covered questions about equipment and shopping patterns, as well as cooking methods and terms, menu planning, nutrition and ingredients.

In the 1980s and 1990s there was considerable change. Two different qualifications were available from the 1990s – home economics and D&T food technology. Although both are described here, D&T food technology had a far greater number of candidates – in this awarding body in 1997 34,067 students took food technology and 25,047 home economics, in 2006 the figures were 20,935 and 3,261 respectively. These figures not only illustrate the decline of home economics by comparison with food technology, but also the very significant decline in numbers overall between 1997 and 2006.

- In home economics, a wider variety of types of questions were introduced to the written papers. Although short answer questions continued to feature in the first section of the paper, they were augmented by multiple choice questions. A section of the paper devoted to data response questions (of which two were presented and one had to be answered) was introduced, and also a section comprising free response questions (again, candidates had to answer one from a choice of two). The practical examination changed from a timed session cooking essentially domestic recipes, to a set of investigations where candidates were required to explore theoretically a 'food based problem', before cooking a number of dishes related to the problem.
- In D&T food technology in the 1990s candidates had to complete a written paper on core D&T content (not related to food). A second written paper assessed the food technology element of the paper and had to complete a piece of coursework for both core content and food content. By 2007 this had evolved to two written papers, both on food content and a coursework component which required

the design, investigation, creation and evaluation of a food product which was suitable for mass marketing. As well as producing the product itself, candidates were required to consider packaging and labelling, as well as target market.

The topic areas covered on the written papers of both the home economics and food technology examinations have broadened from those seen in the O level era, incorporating questions on manufacturing processes, marketing, packaging and labelling, as well as those topics seen in the past, such as nutrition.

Tiering was not applied to this subject by this awarding body until 1997, when the relatively new food technology specification had three tiers for the written paper: Basic (grades G–C), Standard (grades E to A) and Higher (grades D–A*). The home economics examination in 1997 was not tiered. In 2007 two tiers were in place for the written paper of both food technology and home economics examinations.

Implications for the future

The review of cookery qualifications over the years indicates several very stable eras when the qualifications continued in the same format for several decades. There is also clear evidence of how and when changes were made to the way in which the subject was assessed. The current concern about the teaching of cookery in schools centres upon the allegation that students today do not have the skills necessary to create nutritious balanced meals from fresh ingredients in a domestic context. Reviewing the evolution of GCSE and predecessor qualifications does not prove whether this is the case or not, but it does enable us to contextualise the allegation, and assess broadly how, within the context of assessment at 16+, the subject has changed.

It can clearly be seen that cookery qualifications at age 16 have changed over the years to reflect changing social trends in provision of food in the home. For example, written examinations in the UK contain more questions about dietary needs, and fewer asking students to describe 'how to make' a particular recipe, and coursework consists of food based 'problems' often focussed upon a single ingredient, or nutritional need. Ultimately, however, each era has reflected social tendencies of the time, and the manufacturing element of the later era, which forms a large part of the food technology examination, has been in keeping with a society which uses processed food frequently in everyday life.

References

Attar, D. (1990). *Wasting Girls' Time: The history and politics of Home Economics*. London: Virago Press.

BBC News online (September 2005). *Junk food to be banned in schools*. http://news.bbc.co.uk/1/hi/education/4287712.stm

Channel Four Television (2005). *Jamie's school dinners*. http://www.channel4.com/life/microsites/J/jamies_school_dinners/, accessed on 6 November 2007.

Children's Food Campaign (2006). *Response to the Consultation on the Secondary Curriculum Review*. http://www.allianceforchildhood.org.uk/fileadmin/templates/2006/uploads/CFCsecondarycurriculumresponse.pdf, accessed on November 7th 2007.

Christian-Carter, J. (1985). A Brave New World. *Times Educational Supplement*, 19 April 1985.

Cumming, C., Foley, R., Long, A. & Turner, E. (1985). *Where does the proof lie? An account of the Assessment in Home Economics Research Project*. Edinburgh: Moray House College of Education.

DATA (2005). *The Design and Technology Association's views on the KS3 review*. The Design and Technology Association. November, 2005.

DATA (2007). *Briefing paper for members on Secondary Food Education*. The Design and Technology Association. http://web.data.org.uk/data/docs/briefing_dt_assoc_members.doc, accessed on 8 October 2007.

DCSF (2008). *Compulsory Cooking Lessons for all pupils*. Press Notice database, 22 January 2008. http://www.dfes.gov.uk/pns/DisplayPN.cgi?pn_id=2008_0015, accessed on 3 March 2008.

HM Government (2005). *Every Child Matters: Change for Children*. http://www.everychildmatters.gov.uk/_files/F9E3F941DC8D4580539EE4C743E9371D.pdf, accessed on 25 October 2007.

Horne, S. & Kerr, K. (2003). Equipping youth for the 21st century. The application of TOWS analysis to a school subject. *Journal of Nonprofit & Public Sector Marketing*, **11**, 2, March 2003.

McBeth, J. (2005). Children who can't cook... can't sew... can't save. *The Scotsman*, 8 January 2005. http://news.scotsman.com/uk.cfm?id=21112005, accessed on 24 September 2007.

Purvis, A. (2007). *Who is teaching our children to cook?* Waitrose. http://www.waitrose.com/food/celebritiesandarticles/foodissues/9906076.aspx.

QCA (2007). *Design and Technology: Programme of Study, KS3*. http://www.qca.org.uk/qca_12209.aspx, accessed on 9 October 2007.

Raban, S. (2008). *Examining the world. A history of the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Critical Thinking – a tangible construct?

Beth Black Research Division

First published in Research Matters, Issue 3, January 2007

> *Are some outcomes of education too intangible to be measured? No doubt, there are some that we speak of often, like critical thinking.., that [is] so difficult to define satisfactorily that we have given up trying to define [it] specifically. To this extent, they are intangible [and] hard to measure.* (Ebel, 1965)

Forty years on from Ebel's quote, the testing of Critical Thinking has become a flourishing area. In the UK, tests which incorporate a Critical Thinking element include the BioMedical Admissions Test (BMAT), Thinking Skills Assessment (TSA), UniTest, UK Clinical Schools Admissions Test (UKCAT) and Watson Glaser Critical Thinking Appraisal UK (WGCTA-UK). Frequently the stated purpose of these tests is to help Higher Education establishments make admissions decisions, a situation with much precedent in the US where the Law Schools Admissions Test (LSAT) and Medical Colleges Admissions Test (MCAT) are de rigueur for applicants. It seems that to think critically is considered an advantageous or even essential ability for university students on some courses.

But what is Critical Thinking? Is Ebel's pessimistic view now outdated? This article hopes to introduce some of the debates within the construct of Critical Thinking and some of the implications for assessment of Critical Thinking. There are a number of protagonists within the field, and their definitions of what constitutes the construct of Critical Thinking vary enormously: 'chaos at the core' as Benderson wrote in 1990.

The early work of Robert H. Ennis, University of Illinois, propounded a 'pure skills' approach to Critical Thinking. Critical Thinking was defined as 'the correct assessing of statements' (Ennis quoted in Siegel, 1988) and was appended by a list of aspects of statement assessment and criteria. The caveat to this long list is that a complete set of criteria for Critical Thinking cannot be established, that 'intelligent judgement' is also required.

Thus, there are no clear boundaries defining the outer limits of what constitutes Critical Thinking. The implication of Ennis' early position (the 'pure skills' approach), is that if you can pass a test in Critical Thinking, you have Critical Thinking skills. The weakness in this definition is that someone may possess such skills and yet never use them. To *be* a critical thinker and not just be *able* to be one should be an important aspect of the definition. Ennis' (1996) later definition, 'Critical Thinking is reasonable, reflective thinking that is focused on deciding what to believe or do', introduces decision-making into the concept and the idea that Critical Thinking should affect a critical thinker's behaviour, that is, Critical Thinking is exercised and is not just pure skills.

Alec Fisher, Director of the Centre for Research in Critical Thinking at the University of East Anglia, insists that it must be a taught skill, and one that is transferable to other subject domains. He claims an important aspect is metacognition, that is, thinking about one's thinking. Arguably, metacognition can only be achieved through some conscious effort by reference to a good model of thinking. This is where the *teaching* of Critical Thinking comes into play. Additionally, Fisher argues that a critical thinker should exercise and apply these Critical Thinking skills not just in academic studies but in many situations (where appropriate). His definition is:

> *Critical Thinking is skilled and active interpretation and evaluation of observations and communications, information and argumentation.* (Fisher and Scriven, 1997)

Richard Paul, founder and director of Sonoma State University's Centre for Critical Thinking, argues that Critical Thinking courses often teach 'weak-sense' Critical Thinking, where the concepts within can become so atomistic that they are no longer Critical Thinking (just a series of 'moves'). Paul (1992) advocates Critical Thinking in a 'strong' sense. Critical thinkers should look at 'argument networks' or 'world views' and not merely reject an argument network on the basis of an atomistic flaw. One's deepest beliefs and ethical, moral and socio-cultural standpoints should be subject to Critical Thinking. Thus in order to think critically, one must use these skills on oneself; it is a reflective process.

> *Critical Thinking is disciplined, self-directed thinking which exemplifies the perfections of thinking appropriate to a particular mode or domain of thinking.*

John McPeck (1981) of the University of Western Ontario suggests that it cannot be taught as a standalone subject – one is always thinking about something – so that in theory one might offer Critical Thinking for Physics, or Critical Thinking for Geography.

> *In isolation from a particular subject, the phrase "Critical Thinking" neither refers to nor denotes any particular skill. It follows from this that it makes no sense to talk about Critical Thinking as a distinct subject and that it therefore cannot be profitably taught as such. [Critical Thinking] ... is both conceptually and practically empty.*

In short, the construct of Critical Thinking is not precisely defined, nor is it the case that there is a single agreed definition.

Some of this division stems from the experts' fields (though all of the above are involved with the informal logic movement). Those from a philosophical background are interested in employing the tools of logic and reasoning in order to illuminate fundamental truths (with a tradition of more than 2,000 years of reasoning and argumentation). Meanwhile, those from a psychological background, for example, Sternberg and Halpern, are concerned with the thinking process and problem solving rather than logical reasoning. This tradition has evolved not from philosophical argument and discourse, but through experimentation on real subjects. Thus, psychologists may view the philosophers as giving an account of some 'ideal' Critical Thinking abilities, rather than actual performance where limiting factors (e.g. time, information, working memory capacity, motivation) come into play. There are differences between rules of logic and rules of thought. So, psychologists have been concerned with characterising Critical Thinking as it is performed under the limitations of the person and the context or environment. This notion is reflected in the definition of Professor Robert Sternberg (1986) of Yale University:

Critical Thinking comprises the mental processes, strategies, and representations people use to solve problems, make decisions, and learn new concepts.

Thus, one expects from psychologists a more *descriptive* account of Critical Thinking, rather than an *aspirational* account.

Psychologists' definitions and taxonomies of Critical Thinking tend to emphasise problem solving rather than logic. Sternberg's psychological taxonomy of Critical Thinking skills involves metacomponents (e.g. formulating a strategy, monitoring progress in solving a problem), performance components (e.g. inductive and deductive reasoning, spatial visualisation) and knowledge-acquisition components (e.g. encoding and organising information). Interestingly, Critical Thinking tests which stem out of the cognitive tradition do not always separate out Critical Thinking from intelligence (e.g. Sternberg's Triarchic Test of Intellectual Skills).

Unsurprisingly, representatives from each tradition counter attack. Paul (quoted in Benderson) rejects the psychological account on the basis that the puzzles posed by psychologists as critical thinking teaching aids are self-contained or 'monological', that is, are simplistic in that they have a single correct answer and involve adopting just one frame of reference ('weak sense' Critical Thinking). 'True' Critical Thinking should involve 'multilogical' problems, involving multiple frames of reference or argument networks with no single correct answer; only then can a student reflect upon and evaluate their own beliefs. However, Sigel, an ETS researcher notes that 'Philosophers tend not to be empiricists... they just use themselves as sources of authorities. The psychologist is an empiricist who wants to create data that educators can then validate with their own experience.' (quoted in Benderson 1990)

Is there any definition to which the majority of experts would subscribe? Possibly the definition derived from a Delphi study[1] conducted in the United States by Facione (1990). In this study, 46 Critical Thinking experts, consisting of 24 panellists associated with philosophy (including Ennis and Paul), 9 associated with the social sciences, 2 with physical sciences and 10 with education formed a consensus on many aspects of Critical Thinking, including a definition and list of critical skills. The definition, quoted in full, reads as follows:

We understand Critical Thinking to be purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based. CT [sic] is essential as a tool of inquiry. As such, CT is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon. The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgements, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit. Thus, educating good critical thinkers means working toward this ideal. It combines developing CT skills with nurturing those dispositions which consistently yield useful insights and which are the basis of a rational and democratic society.

It is worth noting that this definition has two dimensions to it: cognitive skills and affective dispositions. Facione also provides a detailed taxonomy of skills and sub-skills which helps to define the outer limits of Critical Thinking. However, some commentators regard the list as over-inclusive especially with regard to affective dispositions. Fisher and Scriven (1997) comment that the work is flawed in defining the Critical Thinker rather than Critical Thinking. Certainly, cognitive skills are more readily assessed than affective dispositions in traditionally styled examinations but perhaps, logically, if one wanted to assess the degree to which someone is a Critical Thinker, a personality test would be more appropriate?

Some issues in Critical Thinking literature regarding the construct and their implications for pedagogy and assessment

Thinking which is *not* Critical Thinking?

The corollary to disagreement about what *is* Critical Thinking, is differences of opinion concerning what *isn't*. There tend not to be clearly defined outer-edges of the construct. The Facione Delphi study gives some clues:

Not every useful cognitive process should be thought of as CT. Not every valuable thinking skill is [a] CT skill. CT is one among a family of closely related forms of higher-order thinking, along with, for example, problem solving, decision making and creative thinking. The complex relationships among the forms of higher-order thinking have yet to be examined satisfactorily.

It may matter less to Critical Thinking teachers than to Critical Thinking test-writers as to what defines the outer limits of the discipline. Test-writers face criticisms of construct validity, for example, that their test is really testing the candidates' ideology, common or background knowledge, intelligence or creative thinking rather than, for example, inference, analysis or interpretation skills.

Critical Thinking pedagogy: separate or infused?

Not only is there some lack of clarity in the literature over what to include within a Critical Thinking curriculum, there is also some inconsistency concerning how the curriculum should be constructed. Is Critical Thinking:

(a) something which should be taught as a separate discipline, or

(b) something which is embedded or infused, either implicitly or explicitly, within other subject domains?

Whilst all Critical Thinking protagonists support the view that Critical Thinking should be part of students' educational experience, the conflict is whether its provision should be embedded in subject domains or stand alone as a separate academic discipline. Certainly, McPeck (1981) would, if anything, support the former view, asserting that:

To the extent that Critical Thinking is not about a specific subject, X, it is both conceptually and practically empty. The statement "I teach Critical Thinking", simpliciter, *is vacuous because there is no generalised skill properly called Critical Thinking.*

1. Briefly, the Delphi Method involves the formation of a panel of experts, who participate in a number of rounds of questions involving them sharing opinions. The experts can reconsider them in the light of comments offered by other experts. The overall agenda is to move towards a position of consensus (if not unanimity) on a particular subject.

However, this conflicts with the view of Fisher (2001):

> *Increasingly, educators have come to doubt the effectiveness of teaching 'thinking skills' in this way [implicitly] because most students simply do not pick up the thinking skills in question. The result is that many teachers have become interested in teaching these skills directly…taught in a way that expressly aims to facilitate their transfer to other subjects and other contexts.*

Is Critical Thinking an explicitly teachable skill or a natural disposition?

> *Most of us would claim that we can teach critical thinking, but not be too sure about whether we can change someone's personality.*
> (Fisher and Scriven, 1997)

Whilst some definitions promote Critical Thinking as an explicitly teachable skill, others make more of dispositions. For instance, Ennis's early view of Critical Thinking advocated a 'pure skills' approach, while his later work advocates a 'skills plus tendencies' position. One such tendency involves 'open-mindedness' (Ennis, 2002). As a synonym for openness, this is included as one of the five traits in the so-called 'Big 5' or Five Factor Theory of Personality (McCrae and Costa, 1996) and is widely accepted as a broad personality trait, which many view as fixed in amount or stable throughout adulthood.

McPeck's definition, 'the propensity and skill to engage in an activity with reflective skepticism' (1981), implies another disposition, akin to a 'spirit of inquiry', also present in the definitions advocated by Perkins, Jay and Tishman (1993) in their article aptly entitled 'Beyond abilities: a dispositional theory of thinking'. Interestingly, some critical thinking tendencies (e.g. open-mindedness, being questioning, observant) have some convergence with Guy Claxton's Positive Learning Dispositions (2006), that which a 'capacity to learn' comprises. Despite the use of the term 'disposition', his view is that developing (or teaching) dispositions is a fruitful endeavour. He deliberately clarifies his view of a disposition as 'merely an ability that you are actually disposed to make use of.'

Whether Critical Thinking is an explicitly teachable skill or a (fixed) natural disposition is a pertinent question, both for Critical Thinking teachers as well as people who devise and test Critical Thinking. Equally, what are the valid inferences end users might make from a score or mark obtained on a Critical Thinking Test? Assuming that one can infer that candidate Z has X amount of the ability at the moment of testing, the question is whether one believes this indicates a permanent or transient measure of that person as a Critical Thinker.

Conclusions

So, does Ebel's appraisal of Critical Thinking still hold true forty years on? Far from giving up, there has been considerable endeavour to define Critical Thinking. These attempts have certainly made the concept increasingly tangible and easier to measure, although there is still some way to go before a single definition is accepted by all. Furthermore, the introduction into the arena of over 20,000 students in about 1,000 educational institutions wishing to have their achievement in Critical Thinking certificated has added an additional dimension to Ebel's 'hard to measure' statement. Ebel was undoubtedly right – Critical Thinking is difficult to define satisfactorily and hard to measure. But we have not given up trying.

References

Benderson, A (1990). *Focus. Critical Thinking: Critical Issues*. Princeton: Educational Testing Service.

Claxton, G. (2006). *Expanding the Capacity to Learn: A new end for education?* Opening Keynote address at British Educational Research Association Annual Conference, September 2006.

Ebel, R.L. (1965). *Measuring Educational Achievement*. New Jersey: Prentice Hall.

Ennis, R.H. A concept of Critical Thinking. Quoted in Siegel, H. (1988). *Educating Reason: Rationality, Critical Thinking and Education.* London: Routledge.

Ennis, R.H. (1996). *Critical Thinking*. New York: Prentice Hall.

Ennis, R (2002). http://faculty.ed.uiuc.edu/rhennis/SSConcCTApr3.html accessed on 06/06/2006.

Facione, P.A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction: Executive Summary, The Delphi Report*. Millbrae, CA: California Academic Press.

Fisher, A. (2001). *Critical Thinking: An Introduction*. Cambridge: Cambridge University Press.

Fisher, A. & Scriven, M. (1997). *Critical Thinking: Its definition and assessment*. Norwich: Centre for Research in Critical Thinking.

McCrae, R.R. & Costa, P.T. Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives.* 51–87. New York: Guilford.

McPeck, J.E. (1981). *Critical Thinking and Education*. Oxford, UK: Martin Robertson.

Paul, R. (1992). Critical Thinking: What, Why and How? *New Directions for Community Colleges*, **20**, 1, Spring 1992.

Perkins, D.N., Jay, E. & Tishman, S. (1993). Beyond Abilities: a dispositional theory of thinking. *Merril Palmer Quarterly*, **39**, 1–21.

Sternberg, R.J. (1986). *Critical Thinking: Its Nature, Measurement and Improvement*. ERIC Document Reproduction Service ED272882.

Extended essay marking on screen: Does marking mode influence marking outcomes and processes?

Hannah Shiell, Martin Johnson, Rebecca Hopkin, Rita Nádas and John Bell Research Division

First published in Research Matters, Issue 11, January 2011

Introduction

Technological developments are impacting upon UK assessment practices in many ways. For awarding bodies, a key example of such impact is the ongoing shift towards examiners marking digitally scanned copies of examination scripts on screen rather than the original paper documents. This digital shift affords opportunities to manage and distribute information in ways that are not possible in paper-based marking systems, and this has important quality assurance benefits.

At the same time, however, the shift towards marking scripts on screen has prompted questions about whether the mode of marking might influence the outcomes of the marking process, particularly in relation to essay responses.

Research into comparisons between how people read texts on paper and computer screen suggests that the medium in which a text is read might influence the way that a reader comprehends that text. This is because some of the reading behaviours that support comprehension building, such as seamless navigation and annotation of text, are not easily replicated on screen (Dillon, 1994; Marshall and Bly, 2005; O'Hara and Sellen, 1997; Piolat, Roussey and Thunin, 1997; Rose, 2010).

Additional research also suggests that reading long texts can be more cognitively demanding on screen (Wästlund, Reinikka, Norlander and Archer, 2005), and that this extra demand can have a detrimental effect on how readers comprehend longer texts (Just and Carpenter,1987; Mayes, Sims and Koonce, 2001). In the context of examination marking, there might be concerns that such a mode-related effect might lead to essays being marked less accurately when marked on screen compared with when they are marked on paper.

The theoretical basis for concerns about mode-related influences on essay marking can be summarised by the model presented in Figure 1. This model outlines the potential relationships that are involved when an examiner reads an essay in order to mark it. In summary, literature underpinning the model infers that the shift from marking essays on paper to marking them on screen might be expected to impact upon examiners' manual and cognitive marking processes. This could, in turn, result in examiners having a weaker comprehension of essays when marking them on screen and this might be reflected in the final marking outcome.

Research in this area is therefore a principal concern for awarding bodies and stakeholders, posing potential implications in terms of both the defensibility of assessment outcomes and public trust in the assessment system.

In response to these concerns, researchers at Cambridge Assessment and elsewhere have been investigating how transition from paper-based to screen-based essay marking might influence examiners' marking behaviours and their marking accuracy. Four recent studies have investigated how mode might affect essay marking (Johnson and Nádas, 2009; Coniam, 2009; Fowles, 2008; Shaw and Imam, 2008). These studies,

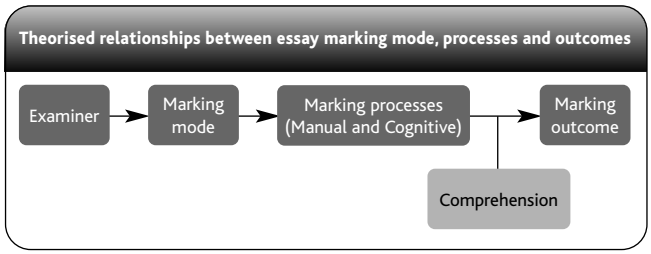


Figure 1 : Theorised relationships between essay marking mode, processes and outcomes

which consider essays of 150 to 600 words, report a negligible mode-related effect on marking accuracy; suggesting little cause for concern as the marking of digital essay images on screen replaces the marking of hard-copy paper essays.

Among the four studies, Johnson and Nádas (2009) is noteworthy as it employs a wider variety of quantitative and qualitative methods. The aim of the project was to broaden investigation beyond the singular consideration of the effects of mode on marking accuracy; to also explore mode-related influences on recognition of essay quality and examiners' marking processes.

As reported in Issue 8 of this journal, the findings of Johnson and Nádas (2009) showed that marking GCSE English Literature essays on screen had no significant effect on marker accuracy when compared with how they were marked on paper, although the examiners did exhibit different marking behaviours when marking in each mode.

The examiners in the Johnson and Nádas (2009) project also experienced significantly heightened cognitive workload levels while they marked on screen. The authors concluded that the examiners may have attained similar levels of accuracy across modes because they had sufficient spare cognitive capacity to accommodate the additional cognitive workload exacted by the screen marking task. Based on this conclusion, the authors suggested that the marking accuracy findings may not generalise to extended essays, therefore recommending that further research should explore mode-related effects in the marking of essays with lengths greater than those which were the focus of the earlier studies.

Research questions and research design

To investigate further the potential links between marking mode and the outcomes and processes of extended essay marking, the current project replicated the Johnson and Nádas (2009) project, replacing GCSE essays with longer Advanced GCE essays.

The current project considered six research questions in three broad areas of enquiry, exploring mode-related influences on (i) marking outcomes, (ii) manual marking processes and (iii) cognitive marking processes. The six questions are displayed in Figure 2.

Research Questions

Mode-related influences on marking outcomes were considered through two research questions (RQs):

RQ1: Is examiner marking accuracy influenced by marking mode?

RQ2: Is examiner recognition of essay quality influenced by marking mode?

Mode-related influences on manual marking processes were considered through three research questions:

RQ3: Is examiner manual interaction (i.e. physical contact) with essays influenced by marking mode?

RQ4: Is examiner essay navigation influenced by marking mode?

RQ5: Is examiner annotation practice influenced by marking mode?

Mode-related influences on cognitive marking processes were considered through one research question:

RQ6: Is examiner cognitive workload influenced by marking mode?

Figure 2 : Research questions

This project used an essay question with a maximum of 60 marks available from an Advanced GCE History unit. One hundred and eighty essays from the June 2009 examination session were selected and split into two samples of 90 essays which were broadly similar in terms of mean marks (from the live session) and mark distributions. Table 1 shows the sample features of the essays used in the current project, compared to the sample used in the Johnson and Nádas (2009) project, which used GCSE English Literature essays.

Table 1: GCE History and GCSE English Literature essay sample features

	N	*Written A4 pages*	*Written lines*	*Estimated word count*
		Mean	*Mean*	*Mean*
GCE History project	180	5.3	123.5	900
GCSE English Literature project	180	3.4	75.8	573

The 180 essays were blind marked on paper by the examination's Principal Examiner (PE) to establish a project reference mark for each essay. A sample of 12 Advanced GCE examiners participated in the project. The examiners were all relatively experienced, holding between 6 and 31 total years' experience (mean 16.8 years) of marking for large-scale educational assessment agencies in the UK. Five of the examiners had some previous experience of marking essays on screen.

The 12 examiners marked one of the two samples on paper and the other sample on screen. To control for essay sample and for marking order, a crossover research design was used and the examiners were randomly allocated to one of four examiner marking groups. Table 2 shows the crossover research design used.

Table 2: Examiner marking groups and essay allocation design

Examiner marking group	*1st marking*	*2nd marking*
1	Sample 1 – Paper	Sample 2 – Screen
2	Sample 2 – Paper	Sample 1 – Screen
3	Sample 1 – Screen	Sample 2 – Paper
4	Sample 2 – Screen	Sample 1 – Paper

Prior to marking, all 12 examiners attended a two day meeting to be trained in using the marking software and to standardise their marking in both paper and screen modes. Semi-structured interviews were carried out with each examiner after the marking period had finished, to allow the researchers to probe and check their understanding of the data.

Findings

Mode-related influences on marking outcomes

RQ1: Is examiner marking accuracy influenced by marking mode?

Marking accuracy was defined as the extent of agreement between the examiner marks and the corresponding PE reference marks. Marking accuracy was investigated by considering the differences between the examiners' marks and the reference marks awarded for each essay. These analyses considered two distinct measures of marking accuracy: *absolute*[1] and *actual*[2] mark differences. These measures give an indication of the magnitude and direction of marking accuracy differences between the examiners and the PE for each essay. Descriptive and general linear modelling statistical analyses were then used to investigate whether examiners' marking accuracy was influenced by marking mode.

Table 3 shows descriptive statistics of absolute and actual mark differences between examiner and PE marks by marking mode. Descriptive analyses of absolute mark differences revealed that in both marking modes half of all examiner marks were awarded within five marks of the corresponding PE reference mark. Given the 60-mark range available for the essays, this suggests close equivalence in the overall magnitude of marking accuracy on paper and on screen. Furthermore, a disparity of just 0.08 marks between mean absolute mark differences was identified across modes. Descriptive analyses of actual mark differences add greater depth to this picture. On paper the overall median absolute mark difference was 0 and mean absolute mark difference 0.02, indicating a balance of leniency and severity in marking. In contrast, on screen the overall median absolute mark difference was 1 and mean absolute mark difference 0.47, indicating a very slight tendency towards more lenient marking on screen.

Table 3: Absolute and actual mark differences between examiner and PE marks by marking mode

	Marking mode	
	Paper	*Screen*
N	*1080*	*1067*
Absolute mark difference		
Mean	5.82	5.74
Standard Deviation	4.86	4.45
Median	4.5	5
Actual mark difference		
Mean	0.02	0.47
Standard Deviation	7.59	7.25
Median	0	1

1. The absolute difference between an examiner mark and the corresponding PE reference mark. This measure assigns all differences a positive value, regardless of their direction. Absolute mark differences therefore provide a clear indicator of the *magnitude* of marking accuracy: smaller absolute mark differences represent greater marking accuracy.
2. The actual difference between an examiner mark and the corresponding PE reference mark. This measure assigns a negative value to marks below the reference mark and a positive value to marks above the reference mark. Actual mark differences therefore provide a useful indicator of the *direction* of marking accuracy: negative actual mark differences represent severe marking and positive actual mark differences represent lenient marking.

To enhance the descriptive outcomes, general linear modelling was used to test the statistical significance of any association between marking mode and marking accuracy (Table 4). No statistically significant association between absolute mark differences and marking mode was identified. This reiterated the findings of the descriptive analyses, confirming that there was no statistically significant mode-related difference in the overall magnitude of marking accuracy.

Analyses of actual mark differences suggested a significant association between marking mode and the direction of marking accuracy. Compared to the reference marks, essays marked on screen tended to be marked slightly more leniently than on paper, with screen-marked essays being awarded an average of 0.44 marks more than paper-marked essays. This small difference was statistically significant at the 5% level. Nevertheless, the effect size of this result, another statistical indication of the estimated strength of the relationship, was almost negligible (partial eta squared = 0.002), highlighting an extremely weak association.

Overall, the general linear models found no significant association between marking mode and the magnitude of marking accuracy, and only a small and extremely weak association between marking mode and the direction of marking accuracy. The findings therefore suggest that the examiners were marking with similar accuracy in both marking modes.

Table 4: Results for general linear models of absolute and actual mark differences between examiner and PE marks

		ANCOVA table (N = 2147)					
Variable	*DF*	*Model 1.1: Absolute mark difference*			*Model 1.2: Actual mark difference*		
		Type III SS	*F*	*p*	*Type III SS*	*F*	*p*
Marking mode	1	4.23	0.26	0.61	106.10	4.14	< 0.05
Examiner	11	789.19	4.34	< 0.01	10481.91	37.20	< 0.01
Essay sample	1	61.07	3.70	0.05	3002.49	117.20	< 0.01
Individual essay (nested in essay sample)	1	13453.51	4.57	< 0.01	54497.48	11.95	< 0.01
Error	1955	32308.83			50083.57		

ANCOVA, analysis of covariance; DF, degrees of freedom; SS, sum of squares

RQ2: Is examiner recognition of essay quality influenced by marking mode?

To investigate this question the features which the PE felt were contributing to essay quality were elicited using a modified Kelly's Repertory Grid method (Kelly, 1955). The PE then rated each of the sample essays against each of these essay features to generate a measure of quality for each essay. Finally, these measures were added to the marking accuracy general linear models to investigate whether examiner recognition of essay quality was equal across modes.

The marking accuracy findings from RQ1 indicated that, on average, the examiners marked essays with similar accuracy on screen as on paper. It was not possible to know, however, whether the examiners' recognition of essay quality was also similar across modes (for example, were the examiners better on screen at marking lower quality essays but worse at marking higher quality essays?). When a measure of essay quality was added to the marking accuracy models, analyses showed that examiner recognition of essay quality was not influenced by marking mode. In other words, the examiners marked high and low quality essays with equal accuracy on paper and on screen.

Together, the findings of RQs 1 and 2 support the conclusion that the accuracy of the examiners' extended essay marks and their recognition of essay quality are not influenced by marking mode, and that accurate and valid marking of extended essays is feasible on screen.

Mode-related influences on manual marking processes

RQ3: Is examiner physical interaction with essays influenced by marking mode?

Data about how examiners tangibly interacted with the essays in both modes (e.g. how they physically touched the essays) were gathered through direct observation of one examiner from each of the four marking groups and augmented by interview evidence from all 12 examiners. The observed behaviours were:

- Tagging – physically holding a position in a text while looking at another text to relate two things;
- Overlapping pages in the line of vision;
- Dynamic Tracking – horizontal physical movement with a finger, pencil or mouse during reading;
- Static Tracking – vertical physical movement with a finger, pencil or mouse during reading;
- Pointing/Circling with a focus on one particular aspect (for example, a word) in the text.

The behaviour profiles gathered for the four observed examiners varied in terms of the number and variety of physical interactions that they used on paper and on screen, suggesting that these behaviours reflect highly personalised reading styles.

Overall, the four observed examiners physically interacted less with the essays on screen. Observation evidence suggested that examiners demonstrated fewer focused attention behaviours (i.e. indications that the examiner was attending to a particular word or piece of information; static and dynamic tracking and pointing/circling) on screen, whilst comparative referencing behaviours (i.e. indications that they were attending to more than one piece of information simultaneously; overlapping and tagging) did not alter across modes.

Some evidence from the examiner research interviews suggested that the increased tendency to interact physically with paper was because it was physically and mentally easier to do so in that mode.

RQ4: Is examiner essay navigation influenced by marking mode?

Data for this area of enquiry were also gathered via direct observation of the four examiners and interview evidence from all 12 examiners. The observations captured data about examiners' navigating behaviours while reading essays in both modes, specifically identifying the number of backward reading movements and movements of focus to other documents, such as mark schemes, question papers and other marked essays. Figure 3 shows the mean number of navigating behaviours per observed page by marking mode.

The observation evidence shown in Figure 3 suggests that examiners attended to the mark scheme, question paper or to other marked scripts relatively infrequently whilst marking, with no notable mode-related differences.

In contrast to the observation evidence, however, in the interviews six examiners suggested that they tended not to return to previously marked

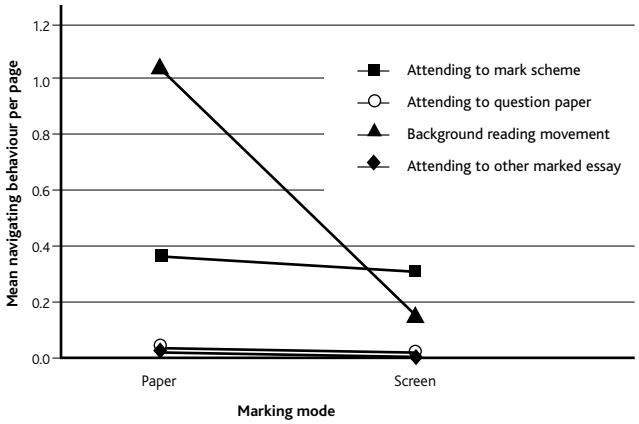


Figure 3 : Mean number of navigating behaviours per observed page by marking mode

essays as readily on screen. Examiners felt that this difference was due to such activity being more difficult to carry out on screen, for example:

> *"Well, I suppose I felt frustrated because it's so difficult…if you wanted to go back three scripts…I thought, 'Oh, can I be bothered with all this clicking and faffing and navigating it, and re-reading it and all this?', and I thought, 'No, I can't'."* (Examiner 8 interview)

Observation evidence also showed that examiners tended to read in a more linear fashion when marking on screen, with fewer iterative or backward reading movements. Examiners suggested in interviews that this was due to the relative difficulty of navigating around essays in this mode:

> *"It's an easier act physically just to turn the page over than to scroll back."* (Examiner 2 interview)

RQ5: Is examiner annotation practice influenced by marking mode?

Thirty essays from essay Sample 1 and 30 matched essays from essay Sample 2 were selected for annotation analysis. The 60 selected essays had each been marked by all 12 examiners and by nature of the research design, each examiner had marked 30 of the selected essays on screen and 30 of the selected essays on paper. Evidence of annotating behaviours was gathered through coded analyses of the marked essays. Again, these data were augmented by interview data from all 12 examiners.

The examiners were able to use a wider variety of annotations on paper than on screen. The screen environment allowed 17 annotation types, including a highlight/underline function. These annotations were built into the marking software following consultation with the examination's PE. For analyses purposes these annotations were termed the 'restricted' annotation palette. Any additional annotations used by examiners when marking on paper were termed the 'extended' annotation palette.

Figure 4 shows the differences in the use of annotations by mode and also by annotation palette. Comparing both the extended and restricted annotation palettes, examiners used an average of 35 annotations per essay on paper and 6 per essay on screen. A Wilcoxon Signed Rank test confirmed that this large mode-related difference was statistically significant ($z = -3.06$, $p < 0.01$, $r = 0.62$). Perhaps this finding is not surprising given that the examiners had access to a limited number of annotation types in the screen marking environment.

When analyses compared only the restricted palette annotations that were available in both marking modes it was found that examiners still annotated less on screen, with a Wilcoxon Signed Rank test confirming this difference to be statistically significant ($z = -2.82$, $p < 0.01$, $r = 0.58$). However, analysis at individual annotation level found that this difference was based on examiners using significantly more underline, sideline and tick annotations on paper. Therefore, when these three annotations were excluded from the overall analysis, there was no significant difference in examiners' use of the remaining restricted palette annotations on paper and on screen.

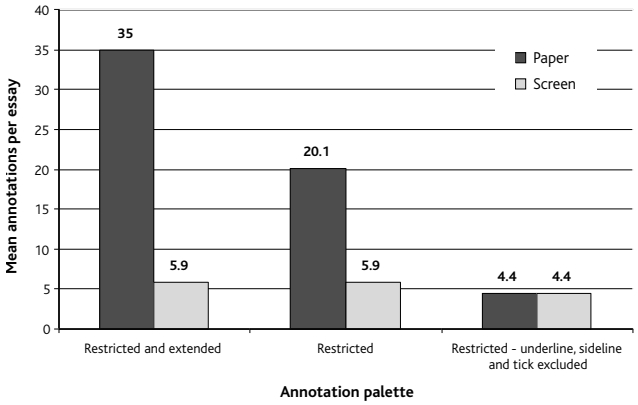


Figure 4 : Mean annotations per essay by marking mode and annotation palette

Examiner interview data were used to help explore the reasons for these mode-related differences. In interviews examiners suggested that they annotated less on screen because the process of using annotations was more difficult and that this might be related to issues of technical usability and their individual levels of proficiency at using the software. Reasons for more limited annotation on screen were also due, in part, to the way that the screen annotation palette sometimes lacked relevance for examiners.

Overall it was evident that physical marking processes were to a large degree idiosyncratic to individual marking behaviours. There was also a clear indication that mode influenced many aspects of examiners' manual marking processes. The physical interaction, navigation, and annotation behaviours that examiners employed for paper-based marking were more difficult for them to replicate when marking on screen.

Mode-related influences on cognitive marking processes

RQ6: Is examiner cognitive workload influenced by marking mode?

Quantitative data about the levels of cognitive workload experienced in each marking mode were gathered using a modified version of the National Aeronautics and Space Administration Task Load Index (NASA TLX) (Hart and Staveland, 1988). The NASA TLX is a self-report survey designed to elicit subjective estimates of the cognitive workload experienced by an individual while performing a specific task. It is underpinned by the assumption that cognitive workload may be represented by a combination of six underlying factors: '*mental demand*', '*physical demand*', '*temporal demand*', '*performance*', '*effort*', and '*frustration*'. The NASA TLX survey was completed twice by 11 of the 12 examiners, midway through their marking sessions in each mode. The survey data enabled a statistical comparison of the cognitive workload

experienced by each examiner across modes to explore whether screen marking was more demanding than paper marking.

Analyses of these data revealed that the examiners experienced greater overall cognitive workload while marking on screen. A Wilcoxon Signed Rank test statistically confirmed that overall cognitive workload was significantly greater on screen ($z = -2.85$, $p < 0.01$, $r = 0.61$). The primary underlying sources of this mode-related difference were identified as the *physical demand* and *fatigue* factors.

Evidence from interview data suggested that the heightened physical demand experienced by the examiners during screen marking was attributed to three key areas of demand: using fine motor skills to operate the computer; maintaining a suitable position at the workstation; and looking at the computer screen. The latter of these physical demands, looking at the computer screen, was highlighted as the most common cause of the fatigue experienced by examiners whilst marking on screen. However, examiner interview comments suggested that this reflected their lack of familiarity with the marking software and might be expected to diminish as their experience of the marking software grows.

Discussion

This project sought to investigate the feasibility of marking extended essays on screen by exploring the potential links between marking mode, essay marking outcomes and marking processes in three broad areas of enquiry;

(i) marking outcomes,

(ii) manual marking processes, and

(iii) cognitive marking processes.

It should be noted that the generalisability of the project findings might be limited by several factors. As a marking simulation exercise, the project differed from a true live marking session in the following key ways:

- The outcomes of the marking exercise had no consequence for candidates, which may have affected examiners' sense of responsibility.
- The marking exercise afforded a comparatively generous time allowance.
- The total marking allocation of 180 essays was comparatively light.
- The previous marking experience of the participating examiners was relatively high.

Marking outcomes

This investigation aimed to consider whether examiners awarded marks which were equally close to the 'true' essay marks in both marking modes. Findings from the statistical analyses suggested that there was no mode-related influence on the magnitude of examiner marking accuracy, but a significant association between marking mode and the direction of examiner marking accuracy was identified. Screen-marked essays were, on average, awarded 0.44 marks more than paper-marked essays. However, the effect size of this result indicated an extremely weak association, and in the context of a 60-mark range the importance of less than half a mark difference is certainly debatable. In light of these perspectives, the findings presented no substantial evidence to indicate that overall marking accuracy was influenced by marking mode.

The examiners' recognition of essay quality across marking modes was also explored. Findings from the statistical analyses suggested that there was no mode-related influence on examiner recognition of essay quality. The examiners attended equally to essay quality when they marked in both marking modes, and the marks awarded recognised that quality.

Together, the marking outcomes findings support the conclusion that the accuracy of the examiners' extended essay marks and their recognition of essay quality are not influenced by marking mode, and that accurate and valid marking of extended essays is feasible on screen.

Manual marking processes

When analyses shifted from marking outcomes to manual marking processes, mode-related influences became more pronounced. The examiners' manual marking processes were broken down into three separate processes: physical interaction, navigation, and annotation. Mode appeared to have an influence on all three of these processes.

The findings show that overall, the examiners physically interacted with essays less on screen than on paper, demonstrating fewer focused attention behaviours when marking on screen. The data did suggest, however, that examiners' physical interaction behaviours were highly personalised, varying widely across individual examiners. Again, when looking at evidence about navigation both within and across essays there were pronounced mode-related tendencies. Evidence showed that the examiners tended to navigate less iteratively on screen and read the essays in a more linear fashion. The most commonly articulated explanation for this difference was the relative difficulty of carrying out traditional paper-based navigation processes on screen.

The examiners in this study also used fewer annotations when marking on screen, due in part to the limited annotation palette available to them on screen. Although the examiners were trained in the use of the software annotation tools it was clear that the examiners still felt that the process of using annotations for marking on screen was too burdensome.

Despite these mode-related differences, examiners were still able to mark extended essays on screen with similar accuracy levels to their paper marking. This implies that the changes in manual marking processes induced by the shift in marking mode did not influence their marking outcomes.

Cognitive marking processes

The examiners experienced greater cognitive workload when marking on screen and this was due to two particular factors – physical demand and fatigue. The examiners attributed the heightened physical demand during on screen marking to the use of fine motor skills to operate the computer, maintaining a suitable position at the workstation or looking at the computer screen. Looking at the computer screen was also highlighted as a common cause of increased and more rapidly arising fatigue.

It is possible that there is an inherent cognitive workload needed when long-held working practices are changed and individuals have to accommodate new ones. The screen marking software influenced examiners' marking processes and these changes could have been initially challenging for the examiners, requiring greater effort. Some of the heightened workload experienced by the examiners could be attributed to their lack of familiarity with the screen marking software, and therefore it is possible that the difference between cognitive workload levels reported across modes might be reduced as examiners' screen marking experience increases.

Conclusion

Returning to the theorised links between extended essay marking mode, processes and outcomes (Figure 1), it appears that mode does have an important influence on some examiner marking processes, but that this does not necessarily influence their marking outcomes. The key practical implication of the findings of this project is that extended essays can be marked on screen without necessarily compromising accuracy. This project supports the conclusions of the Johnson and Nádas (2009) project, and quantitatively demonstrates that the marking of extended essays on screen is feasible. The finding that mode did not present a systematic influence on essay marking outcomes can help to reinforce the defensibility of those marking outcomes and contributes in some way to the maintenance of levels of trust in the assessment system. These findings are of great importance to educational assessment agencies and their stakeholders, and potentially opens the way to the expansion of screen marking to high stakes assessments involving extended essays.

References

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, **15**, 3, 243–263.

Dillon, A. (1994). *Designing usable electronic text*. London: Taylor & Francis.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press, 239–250.

Johnson, M. & Nádas, R. (2009) An investigation into marker reliability and some qualitative aspects of on screen marking. *Research Matters: A Cambridge Assessment Publication*, **8**, 2–7.

Just, M.A. & Carpenter, P.A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn and Bacon.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Marshall, C.C. & Bly, S. (2005). *Turning the page on navigation*. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, 7–11 June 2005, Denver, Colorado, USA, 225–234.

Mayes, D.K., Sims, V.K. & Koonce, J.M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, **28**, 367–378.

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and on-line documents*. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia. ACM Press, New York, 335–342.

Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–589.

Rose, E. (2010). The phenomenology of on-screen reading: university students' lived experiences of digitised text. *British Journal of Education Technology*. DOI: 10.1111/j.1467–8535.2009.01043.x

Shaw, S. & Imam, H. (2008). *On-screen essay marking reliability: towards an understanding of marker assessment behaviour*. Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Wästlund, E., Reinikka, H., Norlander, T. & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: psychological and physiological factors. *Computers in Human Behavior*, **21**, 377–394.

‘Happy birthday to you’; but not if it’s summertime

Tim Oates Assessment Research & Development, **Dr Elizabeth Sykes** Independent Consultant in Cognitive Assessment,
Dr Joanne Emery, John F. Bell and Dr Carmen Vidal Rodeiro Research Division

First published in Research Matters, Issue 8, June 2009

For years, evidence of a birthdate effect has stared out of qualifications data for the United Kingdom; summer-born children appear to be strongly disadvantaged. Whilst those responsible for working on these data have, through mounting concern, periodically tried to bring public attention to this very serious issue, it has been neglected by agencies central to education and training policy. Following a flurry of press interest during 2007 and 2008, it has – justifiably – become a key part of the recommendations which may flow from the Rose Enquiry of the primary curriculum.

Researchers at Cambridge Assessment have had a long interest in the birthdate effect because it is so readily observable in the assessment data that they have worked with (Bell and Daniels, 1990; Massey, Elliott and Ross, 1996; Bell, Massey and Dexter, 1997; Alton and Massey, 1998). More recently, Cambridge Assessment decided to review the issue with the intention to advance the understanding of the extent and causes of the birthdate effect in the English education system (Sykes, Bell and Vidal Rodeiro, 2009). A number of hypotheses have been advanced for its cause – clarity in understanding this fully is a vital part of determining possible remedies. Although the review focuses on understanding the birthdate effect in England, it uses international comparisons as one means of throwing light on key factors.

This article outlines the findings of the review. There is robust evidence from around the world that, on average, the youngest children in their year group at school perform at a lower level than their older classmates (the ‘birthdate effect’). This is a general effect found across large groups of pupils. In the UK, where the school year starts on September 1st, the disadvantage is greatest for children born during the summer months (June, July, August). Individual summer-born pupils may be progressing well, but the strength of the effect for the group as a whole is an issue of very significant concern. Since the effect of being the youngest in the year group holds in other countries where the school year begins at other times in the calendar year, medical/seasonality hypotheses regarding pre-natal exposure to viral infections during the winter months for summer-born children can be ruled out as a major explanation of this effect.

As would be expected, given that one year is a smaller proportion of the total life of a sixteen year old than for a four year old, the birthdate effect is most pronounced during infant and primary school but the magnitude of the effect gradually and continually decreases through Key Stage (KS) 3, 4, and A-level. This pattern is particularly evident in research by the Institute of Fiscal Studies (Crawford, Dearden, and Meghir, 2007). The disadvantage for August-born children over September-born children in attainment dropped from an average of 25% at KS 1 to 12% at KS 2, to 9% at KS 3, to 6% at KS 4 and to 1% at A-level. Despite this decrease, the effect remains significant at GCSE, A-level and in respect of entry into higher education. Likewise, analysis of the results from all of the GCSE examinations taken by over half a million candidates born in England, Wales and Northern Ireland within the same academic year showed a consistent depression in grades achieved for students born from September through to August. In addition, the same pattern of depression was detected in the number of subjects undertaken. Despite decrease in magnitude, the birthdate effect persists until the end of higher education (Alton and Massey, 1998).

Data from 13 LEAs providing GCSE results (undertaken in 1990 to 1994) revealed that birthdate effects were still very evident when all subjects were considered. Summer-borns were the lowest attainers in 10 LEAs and Autumn-born children were the highest attainers in 9 of the Authorities. If gender was included in comparisons then summer-born boys had the greatest disadvantage and autumn-born girls had the greatest advantage. Significantly, it was noted that the difference between these 2 groups was about 1 grade at GCSE in each of 9 subjects taken (Sharp, 1995).

Similarly, the IFS researchers (Crawford, Dearden and Meghir, 2007) found that approximately 6% fewer August-born children reached the expected level of attainment in the three core subjects relative to September-born children (August-born girls 55%; August-born boys 44%; September-born girls 61%; September-born boys 50%). Moon (2003) concludes: ‘If all the pupils in this cohort who were born in the spring or summer terms were to perform at the level of the autumn-born pupils, it would mean that 213 pupils out of a total of 308 improving their GCSE results by an average of 1.5 grades’. The magnitude of the effect has important implications for pupils’ successes and for schools’ overall results.

If the birthdate effect is serious in mainstream education, then it can be argued that it is most serious for those who are struggling in the education system. A disproportionately high percentage of relatively young children in the school year also are referred for special educational needs and many of these appear to be misdiagnosed (Sharp, 1995). The birthdate effect may operate in teachers’ identification of children in need of special education. Teachers may not be making sufficient allowances for the level of attainment against specific curriculum outcomes of the younger members of their classes.

Beyond GCSE, education becomes more selective with choices being made about further participation. Unfortunately, the birthdate effect seems to have serious consequences. The percentage of GCSE students going on to take at least one A-level drops from 35% in September-born students to 30.0% for August-born students (Alton and Massey, 1998). Likewise, September-born students are 20% more likely to go to university than their August-born peers. The Higher Education Funding Council has concluded that ‘…if all English children had the same chance of going to university as those born in September then there would typically be around 12,000 extra young entrants per cohort, increasing young participation by 2 percentage points…’ (HEFCE, 2005).

Given the existence of this effect, it is necessary to identify the underlying cause. There are competing theories regarding birthdate

effects. One is the 'length of schooling' hypothesis – when school admissions are staggered over the year then the youngest have the least schooling. Another is the 'relative age' hypothesis – even with the same length of schooling, the youngest in a year group will be, on average, less mature – cognitively, socially and emotionally – than their older classmates, leading to unequal competition in all three domains that could impact negatively on the younger group. Although it is sometimes difficult to disentangle these two hypotheses, evidence tends to support the latter. Using a common start date does not solve the problem of this type of disadvantage (Daniels, Shorrocks-Taylor and Redfern, 2000).

Teacher expectancy effects may contribute to birthdate effects – teachers may not take children's relative levels of maturity into account when making assessments of their ability and may therefore label younger children as less able than their older peers.

Evidence from developmental psychology suggests that children between the ages of 4 and 5 may not be ready, developmentally, for formal education. Birthdate effects appear to be greatly reduced in countries where formal education begins at a later age. There needs to be a careful consideration of what is best for all children in the early years of schooling, based on solid evidence from psychological research.

The review described here is far more than a simple rehearsal of the findings of a series of relevant studies. It allows an understanding of the accumulation of evidence in respect of the birthdate effect and certain explanations of why it occurs to be discounted. Crucially, the review considers the whole of the education system and this reveals two critical issues. First, that the birthdate effect persists throughout education and training. Secondly, that a strong selection effect may be in operation at all stages – that is, summer-borns are not progressing onto certain routes and into certain levels of education. This effect is not obvious from individual studies limited to specific phases of education. It explains why the summer-borns who get through to the highest level of education are doing well: it is vital to recognise that disproportionately fewer summer-borns actually get to this level *at all*.

Although the existing research is illuminating in respect of the extent of the birthdate effect and of its causes, there is still a need to identify remedies. We believe that work on remedies is not yet sufficiently advanced; substantial, urgent work is required on the means of devising adequate approaches. Although this review was focussed primarily on UK research, it also noted the effect is present in other countries. However, as Bedard and Dhuey(2006) noted, the effect varies from country to country and there is scope for more international work to identify potential solutions to this problem.

From this review, and from the work of comprehensive reviews of the quality of primary and early years education, it is likely that adequate remedy will lie not only in development of a strategy regarding *when* formal schooling should start, but also – at least – in respect of: specific balance in respect of curriculum elements devoted to cognitive, emotional and social development; the training requirements of teaching and support staff; curriculum frameworks; inspection foci; pupil grouping strategy; management of differentiation; and the articulation between early years units and compulsory schooling.

References

Alton, A. & Massey, A. (1998). Date of birth and achievement in GCSE and GCE A level. *Educational Research*, **40**, 1, 105–9.

Bedard, K. & Dhuey, E. (2006). The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. *The Quarterly Journal of Economics*, 2006, **121**, 4, 1437–1472.

Bell J.F. & Daniels S. (1990). Are Summer-born children disadvantaged? The birthdate effect in education. *Oxford Review of Education*, **16**, 1, 67–80.

Bell J.F., Massey A. & Dexter T. (1997). Birthdate and ratings of sporting achievement: analysis of physical education GCSE results. *European Journal of Physical Education*, **2**, 160–166.

Crawford, C., Dearden, L. & Meghir, C. (2007). *When you are born matters: The impact of date of birth on child cognitive outcomes in England*. The Institute of Fiscal Studies: London.

Daniels, S., Shorrocks-Taylor, D. & Redfern, E. (2000). Can starting Summer-born children earlier at infant school improve their National Curriculum results? *Oxford Review of Education*, **26**, 2, 207–20.

Higher Education Funding Council for England (HEFCE) (2005). *Young Participation in Higher Education*. Report Ref. 2005/03. HEFCE, Bristol.

Massey, A., Elliott, G. & Ross, E. (1996). Season of birth, sex and success in GCSE English, mathematics and science: some long-lasting effects from the early years? *Research Papers in Education*, **11**, 2 129–50.

Moon, S. (2003). Birth date and pupil attainment. *Education Today*, **53**, 4, 28–33.

Sharp C. (1995). What's age got to do with it? A study of patterns of school entry and the impact of season of birth on school attainments. *Educational Research*, **37**, 251–265.

Sykes E., Bell J.F. & Vidal Rodeiro, C.L. (2009). *Birthdate Effects: A Review of the Literature from 1990–on*. Research Report. Cambridge Assessment: Cambridge.

All the right letters – just not necessarily in the right order. Spelling errors in a sample of GCSE English scripts

Gill Elliott and Nat Johnson Research Division

First published in Research Matters, Issue 7, January 2009

This article is based on a paper presented at the British Educational Research Association Conference in Edinburgh in September 2008.

Abstract

For the past ten years, Cambridge Assessment has been running a series of investigations into features of GCSE English candidates' writing – the Aspects of Writing study (Massey *et al*., 1996; Massey *et al*., 2005). The studies have sampled a fragment of writing taken from the narrative writing of thirty boys and thirty girls at every grade at GCSE. Features investigated have included the correct and incorrect use of various forms of punctuation, sophistication of vocabulary, non-standard English, sentence types and the frequency of spelling errors. This article provides a more detailed analysis of the nature of the spelling errors identified in the sample of work obtained for the Aspects of Writing project from unit 3 (Literary Heritage and Imaginative Writing) of the 2004 OCR GCSE examination in English. Are there certain types of spelling error which occur more frequently than others? Do particular words occur over and over again? How many errors relate to well-known spelling rules, such as 'i before e except after c'?

Literacy has enjoyed a high profile since 1994 and has been promoted in schools through the introduction of the National Literacy Strategy (NLS). It was unlikely that the 2004 GCSE cohort (the 'population' from whom our writing sample came) was fully exposed to the NLS. This is because many primary schools introduced the NLS from the bottom up, or at least did not implement it for this cohort (in their final year of primary education in the first year of the NLS) on the basis that it would get in the way of preparation for key stage 2 (KS2) national tests (Beverton and English, 2000). This notwithstanding, Beverton and English noted that, in contrast to previous years, grammar was being taught every day and that all teaching staff in the schools observed had a greater awareness of literacy as a subject in its own right. Therefore, the performance of this cohort in spelling is likely to reflect some of the benefits of the NLS.

The study used a stratified random sample of writing taken from a narrative writing task. The only suitable question was found on a paper which formed an alternative to coursework; a question which asked candidates to imagine, rather than to inform, explain, describe, comment, argue or persuade. This option was taken by only 8.3% of candidates – but these amounted to over 5500 candidates from a wide range of schools. The sample was stratified by grade so the fact that this paper was a minority option should be incidental, as the calibre of a candidate achieving a particular grade should be comparable regardless of the route taken through the syllabus. Whilst the possibility existed that schools choosing the examination option might reflect systematic variations in curricular values, comparison of the examination option schools with the entry as a whole did not suggest that the former were unusually socially or educationally selective. The proportions of independent and selective schools as compared with comprehensives and others were the same for the sample as in the overall entry for this English specification.

Spelling errors were identified in the sampled writing by two researchers, working first separately, and then as a team. Each researcher first went through the printed versions of the script samples identifying and counting spelling errors. The two lists of errors and counts were then compared, again grade by grade, and any discrepancies identified and discussed.

The study identified 345 spelling errors in 11,730 words written, and these were reported in Massey *et al*. (2005), with a comparison by grade with samples of writing from 1980, 1993 and 1994. It was shown that a considerable decline in spelling in the early 1990s (compared with 1980) had been halted, and at the lower grades, improved.

Since then, we have conducted a detailed analysis of the 345 misspelled words to see if there is evidence of particular types of error. Each misspelling has been categorised, and five broad types of error identified. These are:

i. sound-based errors,
ii. rules-based errors,
iii. errors of commission, omission and transposition,
iv. writing errors and
v. multiple errors.

This article will present a detailed examination of the misspellings and the process of developing the categorisation system used. A number of words – *woman, were, where, watch(ing), too* and the homophones *there/their* and *knew/new* are identified as being the most frequently misspelled words. Implications for the findings upon teaching and literacy policy are discussed.

Background

The way in which children learn to spell is linked closely to learning to read, and with other elements of learning to write. Westwood (2008) reviewed the literature from 1995 to 2007 pertaining to the strategies used to teach children to read in English in Australia and Great Britain and Wanzek *et al*. (2006) published a review of a large number of intervention studies carried out between 1995 and 2003.

A number of authors have looked at stages by which a child learns to spell. Ehri (1994) identified a 'logographic' stage, whereby a child deduces meaning from the appearance of the words. Later stages include the ability to match letters to speech sounds (Henderson, 1990) and use

these to decode words (read) or to generate their own words (spell). Moats (1995) suggests that a phonetic spelling stage is then attained, with children following a 'one letter spells one sound' strategy. This is the point at which spelling can deviate from conventional 'correct' spellings, especially in English where sound rules do not necessarily match letter rules. At this point the successful speller must memorise specific rules such as grammatical endings, and different words which sound the same but are spelt differently. A study carried out between 1995 and 1998 by the Centre for Language in Primary Education (O'Sullivan and Thomas, 2000) collected data from London primary schools and investigated the teaching and learning of spelling throughout the primary years. Amongst other findings the study reported that it is helpful for teachers to study the mistakes made by individual spellers, in order to assess whether the mistakes they are making are phonetic or visual.

In the UK there have been two main methods of teaching a child to read – synthetic phonics, where children are taught letter sounds before being introduced to whole words (Auger and Briggs, 1992), and analytic phonics, where whole words are introduced from the start. Johnston and Watson (2003, 2004, 2005) have suggested that the reading and spelling skills developed by children taught to read using synthetic phonics are very good.

A number of frameworks already exist which incorporate categories of spelling error. QCA (1999) mentions errors due to unstressed vowels, long 'e', omission of single letters, confusion of consonants and homophones. Homophones are also a feature studied by Hepburn (1991) along with doubling and singling of consonants, articulation, and errors related to inflectional and derivational morphemes. Finally, Mudd (1994) discusses reasonable phonic alternatives – in other words plausible alternative spellings.

Method

The sample of writing from which the spelling errors were identified consisted of the fourth sentence[1] of question 1 (an extended narrative piece of writing) as written by the candidate, and was taken from the scripts of thirty boys and thirty girls at each grade. Where there were insufficient suitable scripts available additional sentences were taken from available scripts. The sentences sampled were keyed into Word™ by a temporary member of staff, preserving all errors of punctuation and spelling. Careful checking was undertaken to ensure that the keying, including errors, had been accurately undertaken. Counts of the numbers of words were then obtained from Word™ software.

Table 1 shows the number of words which were sampled at each grade.

Table 1: Number of words sampled at each grade

Grade	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Number of words	1238	1082	1303	1208	1567	1734	1739	1859

Spelling errors were identified by two members of staff, working first separately, and then as a team. Each person first went through the printed versions of the script samples, grade by grade, identifying and counting spelling errors. The two lists of errors and counts were then compared, again grade by grade, and any discrepancies identified and discussed. At any stage it was also possible to inspect the handwritten scripts to verify the exact marks placed on the paper by the candidate. The benefit of the doubt was given in any case where there was ambiguity, which usually arose as a consequence of either poor handwriting, or poor spacing technique. In some cases it was necessary to look elsewhere in the candidate's script for examples of particular letters or letter combinations, or to look at the spacing between other words to see whether the presence or absence of spacing appeared to be deliberate on the part of the candidate.

1 Everything which appeared between the third and fourth full stop.

Results

Overall numbers of spelling errors

The study identified 345 errors in 11,730 words written. Therefore, 97.1% of words were correctly spelled.

Figure 1 shows the overall numbers of spelling errors by grade. As expected, the number of errors increases by descending grade. Given that spelling errors are one of the (admittedly many) criteria for judging English writing, it would be unexpected if they did not. Figure 2 shows the same data as a percentage of the total number of words, thus adjusting the bars for the number of words written in total (candidates at different grades wrote different numbers of words, and as every word written presents an opportunity for a spelling error, variability in the total number of words might influence the pattern of results). In fact the adjusted graph remains very similar to the raw data.

This paper provides detailed analysis of all the errors to see if it there is evidence for particular types of error. Appendix 1 gives the entire list of words which were spelled wrongly, arranged in alphabetical order.

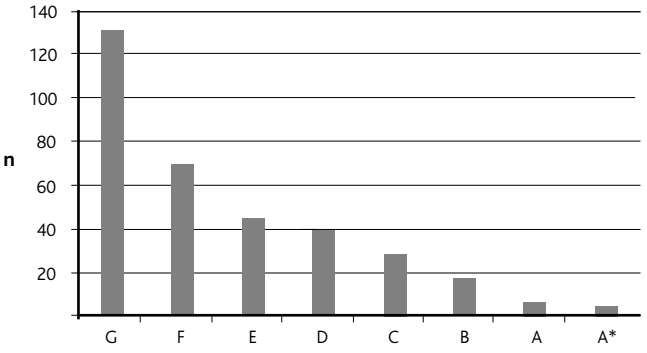


Figure 1: Number of spelling errors by grade

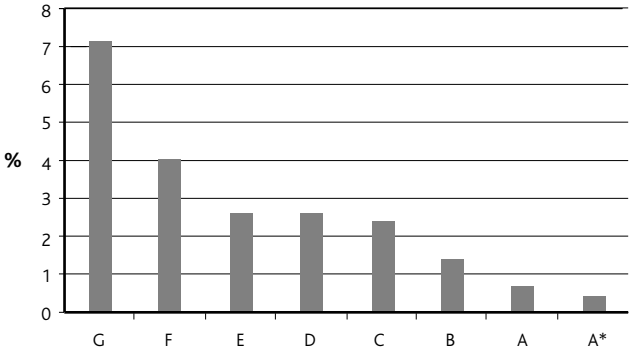


Figure 2: Rate of spelling errors by grade

Frequently occurring misspelt words

A few words occur more frequently than others. Words which appear in the list more than twice are listed in Table 2, along with the frequency of their occurrence, a list of each misspelling and a list of the grades at which the misspellings occur. (The misspellings and corresponding grades are given in the same order, to enable the reader to identify which particular misspelling occurs at which grade.)

Table 2: Frequently occurring misspelt words

Word	*Freq.*	*Misspellings*	*Grade*
before	3	*befor, befor, be for*	GGG
finally	4	*finaly, finily, finaly, finaly*	BGGG
here	3	*he, hear, hire,*	EFG
knew	5	*new, new, new, new, new*	DEEFF
their	3	*ther, there, thire,*	FFF
there	10	*ther, their, their, their, the, their, their, their, their, ther*	BCEEFFGGGG
they	6	*thay, thay, thay, thay, thay, thay*	GGGGGG
too	4	*to, to, to, to*	DEFF
towards	3	*to-wards, to words, to words*	EGG
until	3	*untill, untill, untill*	CCF
watch(ing)	4	*wach, waching, waching, waching*	EFGG
went	3	*when, when, whent*	GGG
were	5	*where, where, where, where, where*	CDDDG
where	4	*were, were, were, were*	DDDF
woman	11	*women, women, women, women, women, women, women, women, women, woneman, women,*	BCDDEEEEFFG
you	3	u, yo, yoy	GGG

Seven of these words – *here*, *their, there*, *too*, *were*, *where*, *you* – appear in published Key Stage 1 lists and *before*, *knew*, *until*, *watch*, *woman* all appear in Key Stage 2 lists.

Although *women* for *woman* is the single most frequently occurring mistake with ten instances (and occurs at every grade from B downwards), the *their*/*there* homophone is a close second, with eight occurrences, seven of which are *there* for *their*.

Misspellings by type

The misspellings presented by candidates have been grouped into broad related categories of error. Categories were derived via a process of grouping together similar error patterns, and are shown in Figure 3. As far as possible the 'types' of error were kept as simple as possible, in the spirit of the original Aspects of Writing (the generic name given to the series of reports produced by Cambridge Assessment, and its predecessor, UCLES) research. This resulted in the following categories:

- Sound-based error – homophones, incorrect consonant, e for y, vowel sound error, morpheme error.
- Rules based error – doubling/singling, text-speak.
- Omission, commission and transposition – single or paired letters added, omitted or transposed.
- Writing error – spacing, end of word missing.
- Multiple errors.

Where a misspelling might fall into several categories (i.e. *accross*, which is both a doubling error and the insertion of an additional letter) the most obvious/most precise error type was allocated; in this case, doubling).

Discussion of error types

Sound-based

Homophones form the first category of error types. 34 of the 345 errors (9.8%) were of this type. The *there/their, know/no* and *knew/new* confusions accounted for nearly half of these. These errors have already been discussed in the section on frequently occurring misspelt words.

Fifteen errors consisted of the transposition of a single wrong consonant. Many of these were phonetically plausible spellings; however, there were instances of a 'k' at the end of *–ing*, instead of the 'g', and of 't' replacing 'd' in *–ed* endings. These were potentially due to articulation error, resulting in spelling error. Two errors involved the transposition of a vowel for a consonant – in both cases 'e' for 'y'.

Fifty-two errors related to the vowel sound. Again (or *agen* according to one such candidate), most of these were phonetically plausible spellings. Nonetheless, many of these words are to be found on the lists of spellings at KS1 and KS2 – e.g. *hospital, heard, some, doctor, they*.

Rules-based

Doubling/singling errors

There were 13 doubling errors and 22 singling errors, together accounting for 10% of all errors. Only one of the errors (*aclimatised*) was an example of an affix error.

Suffix errors

There were 24 suffix errors (7% of the total), of which a very high proportion involved adding *–y* or *–ly* to a word or involved the 'y' to 'i' rule (changing a *y* to an *i* before adding *–ed* (e.g. *replyed*).

Two errors were 'text' (mobile phone/computer text messaging) influenced. Once again these are phonetically plausible alternatives to conventional spelling and are intentionally used in defiance of 'conventional' spelling rules during text messaging. The very small number of these errors was remarked upon in the original report, and it is pleasing to see that candidates seem by and large to be aware that they must not use such devices in a written English examination, however much they are used in social contexts.

Omission/commission of single letter and transposition

Forty-nine errors consisted of the omission of a single letter, whilst thirty-four were the insertion of a single letter. In some cases these were clearly the result of idiosyncratic spellings – notably silent letters. In other cases, the error perhaps owes more to carelessness.

Only ten errors were a straight reversal of two letters, and just one of these related to the 'ie/ei' rule.

Writing errors

Two types of error have been categorised as 'writing' errors. These are errors of spacing – writing two words as one or vice versa, and missing the last letter from a word. In several instances there is evidence from the scripts that candidates did know the correct spelling in the case of the latter category, but had left off the final letter in haste.

Multiple errors

These errors form the arguably most striking type of mistake, and have most effect upon the appearance of the word. First are those misspellings which seem to be made up from two separate errors. For example:

Figure 3: Misspellings by type

Sound-based

Homophones

greatful ruff
hear stairing
here their (x6)
layed there (x2)
lent to (too) (x4)
new (x5) to (two)
no (x2) warn
past weather
piecefully who's
road your

Single consonant confused with another single consonant

ang pup
edje reseption
glanze somethink
looket startet
nothink surport
pankakes trappling
pass warking
pud

e for y

empte lade

Phoneme- grapheme mismatch

agen quiot
clame re-esuring
comfatable saed (x2)
cud sead
deap screeming
devestating secutary
docters suffercated
egere sume
examaning survay
extreamly thay (x6)
frale tomarto
hospitel trough (true)
hourse tumer
hurd uncomfertable
nely (x2) nieve weerdos
paitients women (x10)
parshly wonted
practicle

Rules-based

Suffix rules

ly

definatly slightley
funnyly slowley
highley unnaturely (x2)
luckly

drop e before adding -ing

closeing hopeing
comeing (x4) stareing

y

angery inevitabely
angrey panicy
babys scarey

y to *i*

replyed tryed

ed

answerd offerd

Doubled consonant where should be single

accross ponny
allways pressumably
harrassed quiettly
normall ridding
openned unoccuppied
untill (x3)

Single consonant where should be doubled

acident ofering
aclimatised penciled
asortment poped
atempt siting (x2)
caled spliting
comotion stifly
embarassment stoped (x2)
finaly (x3) sufering
gona tanoy

Text influenced

thanx
u

Omission, commission, & transposition

Single letter omitted

assiting plasic
attemted plesant
belive (x2) quit
complant quitly
coner relised
consious scrunced
contined set
crowed (x2) skatebord
denist stared
drumsicks (started)
easly stiped
emty stroger
enviroment studing
everone subconsiously
exept suprise
exusing suprisingly
fustrating tak
grove tepted
is (his) the (they)
newpaper throt
nuber wach
ofering waching (x2)
overwelming wat
pachy were (where)

Unnecessary letter inserted

alls otheir
anixiously plance
anouther propbable
diden't site (sit)
disrupte smocking
doupe stat
hand (and) throught
has (as) tould
heared verey
hoppe watiching
markers whant
minde whas
minuites whent (x3)
off (of) where (x5)

Extra syllable inserted

partening woneman

2 letters reversed

brian minuets
ect recieve
frist retruned
gentelman thier
minuet wrinkeld

Writing

Spacing/writing two words as one or vice versa

alot to-wards
alright a nother

End of word missing

befor (x2) of
ever on
feminin the (there) (x2)
gonn (gonna) ther (there)
he (her) though
he (here) tong (tongue)
I (it) use (used) (x2)
imagin yo
nam

Multiple

Two 'simple' errors

abound (about) oader
angshuse poedem
babal reapted
behide remmeberd
be for sopose
costrophobic sopted
diesese stir (stare)
diside sufícate
enbarrased sumbleing
finily to words
formiler toke
glome tort
gourges unaturaly
imediatly when (went)
impaitientley (x2)
manegd wittnes
nieghbor

Part of word missing/ severely misspelled

apoched
appment
blacks (blackouts)
canures (cancerous)
he's (here's)
imaging (x2)
prespetion
prest (pressed)
pust (pushed)
registed
scowered
themsefs
trould (trouble)

3 or more mistakes

alla (all of)
handon (handsome)
immeiadtley
solisters
thire (their)

Extreme phonetic errors

ant shaght (anxious)
asct (asked)
corried door (corridor)
faunt (thought)
hast (asked)
nufse (nervous)
or wright (all right)
torck (talk)

impaitientley consists of two separate inserted letters; *impa(i)tientl(e)y*

nieghbor consists of a transposition and an omitted letter (in UK spelling);*n(ie)ghbo(u)r*

Second are those errors where a whole part of a word is either missing or severely misspelled. The third category within this group contains those few words with three or more individual mistakes, and it was one of the misspellings – *immeiadtley* – which prompted the title of this article – all the right letters, just not necessarily in the right order. Finally, there are a group of words which bear little physical resemblance to their correct spellings, yet have clear phonetic links with them. These are referred to as extreme phonetic errors. It is possible that this latter category may be related to the very specific types of error made by people with dyslexia, but further discussion of this is beyond the scope of the present article.

Discussion

This article has attempted to categorise spelling errors made by students in their GCSE English examination in 2004 into various categories. The purpose of the research was to establish whether certain spelling errors – or certain categories of error – are particularly common, and how they relate to spelling conventions, as taught within schools.

The study has identified five categories of spelling error which further subdivide into sixteen subsections. The categories were derived from the errors observed, rather than from existing categories, so there may be other groups of spelling error which have not been discussed here, simply because they were not encountered. In general, most misspellings fall into the first three categories: sound-based error, rules-based error and errors of omission, commission and transposition. The first two of these categories contain many misspellings that are undoubtedly very familiar to teachers. However, there are no particular sub-categories that are particularly prone to more errors in our sample than others. English is a language which has more than its fair share of idiosyncratic spellings and complex spelling rules. Not surprisingly, many of these errors are connected with those. However, within the category of a single additional letter, there were a number of examples of an unnecessary silent 'h' – *where (were)*, *whant, whas*, which are worthy of comment.

The category of omission, commission and transposition is more difficult to interpret. It is quite possible that many of these errors occurred as a result of the examination conditions under which candidates were writing, combined with, perhaps, a lack of effective proof-reading of their final piece. The sub-category of writing errors, where the ends of words are missing, could in some cases be due to the same issues. However, the spacing of two words as one, or vice versa, is almost certainly due to candidates' perceptions of those words. Finally, the category of multiple errors produces words which look least like conventional spellings. Interestingly, two simple errors can produce a word that is almost unrecognisable, and it is important to be able to decode these errors for what they are, rather than simply seeing a very distorted word.

Fifteen individual words were identified as occurring with relatively high frequency. In particular, two of these were seen far more often than others. They were the *there/their* homophone, which has been known to be problematic since time immemorial, and *women* for *woman* (not vice versa). *Knew/new* and *know/now* also occurred with relative frequency, but again, this is unlikely to surprise the teaching profession.

A major limitation to the data presented here is the fact that there is no control over which words candidates choose to use. Therefore the study is not a 'fixed' spelling test, and cannot be generalised in the same way as reports of spelling tests. A word spelt wrongly just once does not mean that 479 students can spell it, simply that they did not necessarily try. It would be possible to investigate correctly spelt words to give the other side of the picture, but that would be an enormous task.

There is clearly no single over-riding type of error which is made by the group of GCSE students from whom we have sampled. Those errors that are made are varied, and although it is disconcerting to note the number of most frequently occurring errors which are taught at Key Stage 1, it is, on the other hand, heartening to see how few (relatively speaking) errors are made, when you consider the number of words written overall, especially given that the text was written under examination conditions with no access to dictionaries.

References

Augur, J. & Briggs, S. (1992). *Hickey Multi-Sensory Language Course*. London: Whurr Publishers Ltd.

Beverton, S. & English, E. (2000). How are schools implementing the National Literacy Strategy? *Curriculum*, **21**, 2, 98–107.

Ehri, L. C. (1994). Development of the ability to read words: update. In: R. Ruddell, M. Ruddell & H. Singer (Eds.), *Theoretical Models and Process of Reading*. Newark, Del.: International Reading Association.

Henderson, E. (1990). *Teaching Spelling*. Boston: Houghton Mifflin.

Hepburn, J. (1991). Spelling categories and strategies. *Reading*, April 1991.

Johnston, R. S. & Watson, J. (2003). *Accelerating Reading and Spelling with Synthetic Phonics: A five year follow up*. *Insight 4*. Edinburgh: Scottish Executive Education Department.

Johnston, R. S. & Watson, J. (2004). Accelerating the development of reading, spelling and phonemic awareness. *Reading and Writing*, **7**, 4, 327–357.

Johnston, R. S. & Watson, J. (2005). *A seven year study of the effects of synthetic phonics teaching on reading and spelling achievement*. *Insight 17*. Edinburgh: Scottish Executive Education Department.

Massey, A. J. & Elliott, G. L. (1996). Aspects of Writing in 16+ English examinations between 1980 and 1994. Occasional Research Paper 1. University of Cambridge Local Examinations Syndicate.

Massey, A. J., Elliott, G. L. & Johnson, N. K. (2005). Variations in aspects of writing in 16+ English examinations between 1980 and 2004: Vocabulary, Spelling, Punctuation, Sentence Structure, Non-Standard English. *Research Matters: A Cambridge Assessment Publication*. Special Issue, November 2005.

Moats, L. C. (1995). *Spelling: Development, Disability and Instruction*. Baltimore: York Press.

Mudd, N. (1994). *Effective Spelling: A practical guide for teachers*. London: Hodder & Stoughton.

O'Sullivan, O. & Thomas, A. (2000). *Understanding Spelling*. London: Routledge.

Qualifications and Curriculum Authority (1999). *Technical Accuracy in written English: Research findings*. London: QCA.

Wanzek, J., Vaughn, S., Wexler, J., Swanson, E. A., Edmonds, M. & Kim, A.H. (2006). A synthesis of spelling and reading interventions and their effects on the spelling outcomes of students with LD. *Journal of Learning Disabilities*, **39**, 6, 528–543.

Westwood, P. (2008). Revisiting issues in spelling instruction: A literature review 1995–2007. *Special Education Perspectives*, **17**, 1, 33–48.

Appendix 1: Alphabetic list of words which were spelled wrongly in a sample of GCSE English writing

A
about
accident
acclimatised
across
again
all of
all
always
a lot
all right
and
angry
another
anxious
anxiously
answered
approached
appointment
as
asked
assortment
assisting
attempt
attempted

B
babble
babies
before
behind
believe
blackouts
brain

C
called
cancerous
claim
closing
coming
comfortable
commotion
complaint
corner
conscious
continued
corridor
claustrophobic
crowded
could

D
deep
decide
definitely
dentist
devastating
didn't
disease
disrupt
doctors
dope
drumsticks

E
eager
easily
edge
embarrassed
embarrassment
empty
environment
etc
every
everyone
examining
except
excusing
extremely

F
face
familiar
feminine
found
finally
first
frail
frustrating
funnily

G
gentleman
glance
gloomy
gonna
gorgeous
grateful
groove

H
had
handsome
harassed
her
hear
heard
here
here's
highly
his
hope
hoping
hospital
horse

I
imagine
imagining
immediately
impatiently
inevitably
it

J
K
knew
know

L
lady
laid
leant
looked
luckily

M
makers
managed
mind
minute
minutes

N
naive
name
nearly
neighbour
nervous
newspaper
normal
nothing
number

O
odour
of
off
offered
offering
one
opened
other
overwhelming

P
pancakes
panicky
partially
parting
patchy
patients
passed
past
peacefully
pencilled
pleasant
place
plastic
podium
pony
popped
practical
prescription
pressed
presumably
probably
pub
pushed

Q
quiet
quietly
quite

R
realised
reassuring
reception
receive
registered
remembered
repeated
replied
returned
riding
rode
rough

S
said
sat
scary
scoured
screaming
scrunched
seat
secretary
sit
sitting
skateboard
slightly
slowly
smoking
solicitors
some
something
splitting
spotted
stare
staring
started
stiffly
stopped
striped
stronger
studying
stumbling
subconsciously
suffering
suffocate
suffocated
support
suppose
surprise
surprisingly
survey

T
talk
tannoy
tempted
thanks
their
themselves
there
they
thought
throat
too
told
took
tomato
tongue
towards
trampling
tried
trouble
true
tumour
two

U
uncomfortable
unnaturally
unoccupied
until
used

V
very

W
walking
want
wanted
was
watch
watching
weirdoes
went
were
what
whether
whose
where
witness
woman
worn
wrinkled

X
Y
you
you're

Cambridge Assessment

Established over 150 years ago, Cambridge Assessment is the University's international exams group, comprising three exam boards as well as the largest educational research capability of its kind. We are a not-for-profit organisation.

Cambridge Assessment plays a leading role in researching, developing and delivering educational assessment to eight million learners in over 160 countries every year. We are an integral part of education and training, advising governments and non-governmental organisations, as well as partnering industry leaders around the world.

Cambridge Assessment's three exam boards are:

- University of Cambridge ESOL Examinations (English for Speakers of Other Languages)
- University of Cambridge International Examinations
- OCR (Oxford Cambridge and RSA Examinations)

University of Cambridge ESOL Examinations is the world's leading range of qualifications for learners and teachers of English. Each year over 1.5 million people in 135 countries take our Cambridge ESOL examinations. Thousands of universities, employers, and government ministries rely on Cambridge ESOL certificates as proof of English language ability.

University of Cambridge International Examinations is the world's largest provider of international qualifications for 14–19 year olds. We offer the Cambridge International Curriculum for 5–19 year olds, professional qualifications for teachers and vocational qualifications for adult learners. We work in partnership with ministries of education, qualifications authorities and examination and assessment boards around the world.

OCR is one of the UK's leading and most respected examining bodies, delivering qualifications and support services to more than 13,000 schools, colleges and other institutions across the country. OCR is recognised as having the most integrated offering of qualifications, covering general, vocationally-related and occupational.

In addition to the three exam boards we have **Admissions Tests and Special Testing** which co-ordinates and manages the development and delivery of new types of products and services, with a primary focus on Admissions Tests for entry to Higher Education, together with Assessment for Learning. These were developed in response to requests from the higher education institutions to help them differentiate between highly able candidates.

We also have a commitment to sharing knowledge and through the **Cambridge Assessment Network** we have established a centre of excellence in assessment for the personal development of those involved in the process of assessment.

www.cambridgeassessment.org.uk

CAMBRIDGE ASSESSMENT

Cambridge Assessment
1 Hills Road
Cambridge CB1 2EU
United Kingdom
tel +44 (0) 1223 553311
fax +44 (0) 1223 460278
www.cambridgeassessment.org.uk