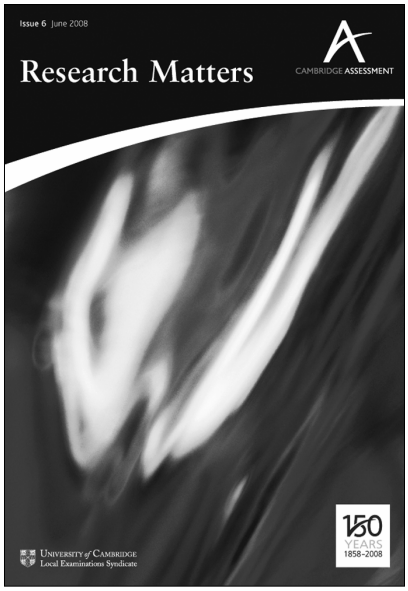
Issue 6 June 2008

Research Matters

CAMBRIDGE ASSESSMENT

150 YEARS 1858–2008

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 : 6

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

A week in politics is a long time. In the light of this, one hundred and fifty years in assessment and qualifications is an eternity. With this timeframe, and with the book 'Examining the world' charting the profound changes in circumstances and structure which Cambridge Assessment has been through, it is perhaps important for current researchers in the organisation to see themselves not as individual investigators but as both the inheritors of a long tradition of enquiry and as custodians and contributors to a continuing bequest to future generations of learners and assessment professionals. Commentators on educational research have bemoaned 'paradigm wars' which have wracked the field, this coupled to concerns over the low levels of genuine accumulation of knowledge – in comparison with other areas of scientific enquiry. By contrast, the analyses of method and the empirical studies described in this edition of *Research Matters* are explicitly designed to add to knowledge accumulation on assessment and qualifications – to build on an established body of operational and research work. The studies place great emphasis on the design of enquiry, and on careful adoption of appropriate method. It builds foundations, we hope, for the next 150 years of robust and useful research.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

In the first article Johnson explores the relationships between, and the importance of, respect, relationships and responsibility in the context of assessment related research. He shares practitioner knowledge and draws from the work of eminent researchers, particularly in the vocational field.

The next four articles focus on the judgements made by examiners and the factors that influence their decisions. Crisp's work draws on a study of the processes involved in marking and grading and investigates which features of student work examiners and teachers attend to and whether these are always appropriate. In his article on marking essays on screen Shaw considers how on screen essay marking affects assessment and marking reliability. His research is carried out in the context of Cambridge International Examinations' (CIE) Checkpoint English Examination. Johnson moves the focus of human judgement into the vocational arena in his article on holistic judgement of portfolios. He considers how assessors integrate and combine different aspects of an holistic performance into a final judgement. Johnson and Shaw discuss another aspect of decision making in their article on annotation, considering the way that assessors build an understanding of textual responses using annotation when marking. They review various themes and models of reading comprehension before considering both the formal and informal influences of the annotation process.

Elliott's article on the examination of cookery from 1937 to 2007 provides interesting information on the way the subject has changed. This is a very topical theme as calls for a return to 'traditional' home cooking has become the subject of much debate. Elliott looks to the past and the present to see how the subject has evolved over the years. Black's article on Critical Thinking looks forward to a growing area of learning and assessment. A number of new Critical Thinking products are in development and Black's work provides coherent guidelines in the form of a definition and taxonomy upon which new developments can be based. Oates looks to the future in his article and considers what lies ahead in the next 150 years. He considers trends in assessment and discusses some of the key issues and challenges facing assessment systems in the years ahead. Roberts highlights some of the activities surrounding Cambridge Assessment's 150th anniversary and provides information about the 34th International Association for Educational Assessment (IAEA) Annual Conference to be hosted in Cambridge in September 2008.

Sylvia Green *Director of Research*

'3 Rs' of assessment research: *Respect, Relationships and Responsibility* – what do they have to do with research methods?

Martin Johnson Research Division

Introduction

This article developed from a speculative email to Dr Helen Colley from the Education and Social Research Institute (ESRI) at Manchester Metropolitan University. I had read one of her conference papers which used a qualitative case study method to explore the interaction of formal and informal attributes of competence-based assessment (later developed into a journal article; Colley and Jarvis, 2007). I wanted to understand how she had gathered some of the rich contextual data in her work which covered a set of social interactions around assessment activities in various vocational settings. Following this initial contact it was clear that there was an overlap between methodological considerations being discussed at ESRI and ideas that were floating around between some members of the Research Division at Cambridge Assessment. These issues centred on the merits and challenges of using qualitative research methods, and how these could contribute positively to the study of assessment. These discussions resulted in the convening of a well-attended research seminar in Cambridge on the 31st October 2007. This seminar, involving Helen and Professor Harry Torrance was called *'How can qualitative research methods inform our view of assessment?'* This article is based on the paper that I delivered at that seminar, with a few additional elements reflecting some of the comments received that afternoon.

The idea for a qualitative methods seminar was prompted by two separate but related issues. The first relates to the Research Division's growing involvement with the wider research literature in the vocational learning field. This literature sometimes draws heavily on qualitative methods to gather rich data about learners and learning conditions in a variety of contexts. An increasing awareness of this vocational literature has also made me more conscious of my own limited understanding of this area of methodology, and so to some extent the seminar grew out of a desire to share research practitioner knowledge and to help to contribute further to the Division's combined research capacity.

The second 'alliterative' prompt for the seminar came from three overlapping themes. The first arose from hearing a lecture given by Randy Bennett at a University of Cambridge International Examinations research conference in 2006 (Bennett, 2005). This paper was then the subject of a response from Tim Oates (Oates, 2007). Finally, another of my recent research projects had led me to pick up a reference to some work by Ann Oakley (Oakley, 2000). I argue that the inter-related strands of the 3Rs of respect, relationships and responsibility that are inherent to these three references can be used to explore some of the issues that influence the instigation and practice of assessment-related research at Cambridge Assessment.

Respect

Randy Bennett argues that research has an important role in reinforcing the integrity of and respect for an organisation as it is perceived by others. He considers the way that non-profit assessment agencies can come to occupy a niche in the educational assessment market place by 'taking on the challenges that for-profit agencies will not, because those challenges are too hard, or investment returns might not be large enough or soon enough' (2003, p.9). An important aspect of this integrity arises from the ability to ask those questions that the other agencies do not. A research division, through its interactions beyond its host organisation and access to outside academic linkages, can view the host organisation from a different perspective to those whose main concern is at an operational level. This gives research an obvious strategic role, enabling researchers to draw upon such perspectives to generate important research questions.

Relationships

Tim Oates (2007) argues that there has been a strong traditional link in the UK between independent assessment agencies, such as Awarding Bodies/Examination Boards, and the communities that they serve. He goes on to point out that this relationship has supported an important accountability function by keeping such agencies responsive to the needs of those that they affect most directly, these principally being the schools and learners with which the agencies interact. Again, I would maintain that research has an important role to play in this interaction through providing evidence of the ways that the practices of our own organisation influence the learning and experiences of others. Here I think it is important to introduce the concept of 'subjective agency' since this is important to the points that follow. Altieri (1994) suggests that subjective agency is an account of human agency in all its dimensions, from psychological through to political, and an important aspect of this agency involves an agent being able to reflect 'self critically'. I argue that this can be translated across to our own 'institutional self', where we can reflect critically on our own position within the wider educational system. This has a number of methodological implications which are discussed later. The key notion of 'subjective agency' also brings us to the third 'R'.

Responsibility

Acknowledging that the activities of our own organisation directly influence the lives of others brings with it responsibilities. Ann Oakley

states that 'the goal of emancipatory social science calls for us to ensure that those who intervene in other people's lives do so with the most benefit and the least harm' (2000, p.3). Oakley's position is to make sure that any activities that are likely to affect others are based on sound research evidence. In our case, understanding impact might involve space for the voices of those affected by educational assessment, and this has obvious implications for the methods chosen to achieve this.

The common strand that unites the three 'R' elements is the conceptual importance of the ability to act 'self-critically' and to understand how an organisation interacts with, influences, and is influenced by, the system within which it operates. So what does this mean for method?

Bourdieu and Wacquant (1992) would suggest that one of the key criticisms of research might be that its practices are limited by its traditions and habits of thought. A key tenet of Bourdieu's theoretical stance is that professional practices are constrained by the structural factors pertaining to their position. He also cautions that any research questions that are being generated could be partial if they only rely on established orthodoxy. This is because these orthodoxies have been connected with the organisation's historic position within the field and thus are unlikely to question conventional perspectives. This places the onus on researchers to first of all recognise the constraints affecting their practice and to constantly question the prevailing techniques. The importance of this final point is made by Oakley. She argues that the historical development of scientific thought has been marked by the presence of some methods that have traditionally only occupied spaces at the edge of the dominant vision. This concept also links to the process of paradigm shift identified by Thomas S. Kuhn to explain how scientific thought develops through the relative capacities of dominant and emerging paradigms to adequately explain different phenomena (Kuhn, 1970).

The notion of 'subjective agency' has important implications for research methods because it is based on assumptions that encourage the use of qualitative research methods. To explain this notion the contested assumptions about the nature of social reality that have dominated a polarised discourse in social science need to be considered. Cohen and Mannion (1994) highlight the way that social science has typically been characterised as having two polarised views of social reality; 'objectivist' and 'subjectivist' (Figure 1). Those who have an 'objectivist' (or positivist) tendency argue that social science mirrors natural science, where a hard, external, objective reality exists with universal laws or constructs waiting to be detected, quantified and measured. This perspective supports the use of controlled experimental methods to analyse the relationships and regularities between selected factors, using predominantly quantitative techniques. This paradigm has been used in one recent Research Division project which investigated whether giving test takers a graded outcome

Figure 1: Social science and 'ways of knowing'

Objectivism/positivism	Subjectivism
• A tangible, external, objective reality exists	• The social world differs from inanimate natural phenomena largely because of our involvement with it
• Methods used to analyse the relationships between selected factors in the world	• 'Reality' is something open to interpretation and is difficult to control
• Tends to involve deductive, quantitative identification and measurement of constructs	• Methods try to understand the ways in which individuals create, interpret and modify the world
	• Tends to involve inductive, qualitative aspects

affected their motivation (Johnson, 2007). The project constructed matched experimental and control groups of test takers, subjected them to different testing conditions, measured their outcomes through a survey method, and analysed these outcomes quantitatively. Whilst this analysis implied a significant relationship between the conditions and outcomes, it also carried within it an inherent frustration that any interpretations being made about why these significances existed could not be any more than weak conjecture.

Polarised discussions about method paradigms are still present within some academic discourses. This is particularly the case in the context of the US where debates about 'scientifically-based research' have followed in the wake of the *No Child Left Behind* agenda (Bliss *et al*., 2004; Maxwell, 2004). Some would argue that arguments that focus on the polarisation of objectivism and subjectivism are less useful than discussions about scientific realism since this provides an opportunity to overcome harmful polarised confrontation and a potential foundation on which to develop research dialogue. House (1991) outlines the scientific realist position. He argues that knowledge is both a social and historical product and that the task of science is to not only invent theories to explain the real world, with its complex layers, but also to test such theories through rational criteria developed within particular disciplines. Furthermore, causalities need to be understood in terms of 'probabilities' and 'tendencies'. This is because behaviour is considered to be a function of agents' basic structures and that events are the outcomes of complex causal configurations.

Discourses of scientific realism also offer the opportunity to overcome potential problems encountered by research. The frustration in the grading and motivation research project reported earlier resonates with some recent concerns expressed by practitioners from the healthcare field. Some clinicians, for example Greenhalgh (1999) and Rapport *et al*. (2004), argue that whilst scientific Randomised Controlled Trial (RCT) methods have been successful in proving the efficacy of particular medical interventions, such methods fail to take account of some of the messy, individualistic, 'irrational' reality that can ultimately affect the success of those treatments. Rapport *et al*. argue that 'only through an appreciation of the integration between human experience and bioscientific treatments of disease, be it within historical, sociological, medical or ethical genres, can we hope to reach clarity of understanding that befits the problem' (2004, p.6). This kind of perspective helps to explain why RCT methods might find it difficult to explain why some individuals just fail to take their medication, which in reality leads to the reduced overall efficacy of such interventions.

Realist discourse implies the need for a wider research paradigm which considers individuals within their own context. What these clinicians argue for is another 'way of knowing' that accommodates a subjectivist outlook. This perspective emphasises that the social world differs from inanimate natural phenomena largely because of our involvement with it, and that 'reality' is something open to interpretation and which is difficult to control. This perspective also suggests that research should focus on the way that individuals construct, interpret and modify the world in which they find themselves. It also suggests that research evidence should take context into consideration since this can be an influence on behaviour. An important consideration is also to reduce the distance between the researcher and the research subject, since shared frames of reference can facilitate the making of legitimate inferences. The complexity inherent in this subjectivist outlook leads to some exciting methodological possibilities.

Questioning the objectivist paradigm in practice can lead to the adoption of mixed qualitative and quantitative techniques. This sort of discussion has already caused a stir in the medical humanities where some have referred to this area of methodology as 'the edgelands' (Rapport *et al*., 2004). They use this metaphor to conjure up the cluttered geographical crossover areas where urban and rural landscapes merge, suggesting that overlapping research paradigms might be similarly messy when they converge. Research beyond the positivist paradigm requires a terrain where new approaches to knowing can be explored. Again, recent work in the Research Division can be characterised by such a metaphor, with one example being the marker annotation project (Crisp and Johnson, 2007). This project used a mixture of a controlled verbal protocol elicitation technique with semi-structured interview and observation methods to gather data about the annotation practices of members of different marking groups. This analysis used a community of practice metaphor to frame an understanding of the patterns within the data, inferring connections between the individuals in the study. A more recent project, the *OCR Nationals* holistic assessment project (Johnson, *in press*), replicated this method but complemented it further by gathering ethnographic observational data of individuals' working in their normal context. This approach then also allowed for the consideration of how value systems might have influenced the behaviour of the participants.

I think the metaphor of 'the edgelands' is very useful for two reasons. First, it implies the need for researchers to consider how methods might be combined to make findings more powerful. Schulenberg (2006), in a paper examining police officers' discretionary decision-making processes with young offenders, argues that mixed methods allow triangulation, complementarity (where findings gained through one method offer insights into other findings) and expansion (of the breadth and scope of the research beyond initial findings). This resonates with the sentiments of Pope and Mays (1995) who also argue that mixed methods can add value to medical evidence gathering because 'qualitative methods can help to reach the parts that other methods cannot reach'. Secondly, I think 'the edgelands' metaphor is very useful because it reminds us that there are areas of activity where we might have a limited understanding and where our efforts need to be directed. One example of this might be in the areas of so called 'non-standard' learning contexts and the learners within them who are affected by educational assessment.

In conclusion, the Research Division has a critical role in supporting the integrity of Cambridge Assessment. Implicit in this is the need to engage in the areas where assessment affects the lives of others. This means not only asking the difficult questions but also having the appropriate methodologies to try to answer them. An important aspect of this entails our continued interaction with other researchers beyond our own institution.

References

Altieri, C. (1994). *Subjective agency: A theory of first-person expressivity and its social implications*. Oxford: Blackwells.

Bennett, R. (2005). *What does it mean to be a nonprofit educational measurement organization in the 21st Century?* Princeton, NJ: ETS.

Bliss, L. B., Stern, M. A. & Park, H. (2004). *Mixed Methods: Surrender in the Paradigm Wars?* American Educational Research Association annual conference, San Diego CA.

Bourdieu, P. & Wacquant, L. (1992). *An invitation to reflexive sociology*. Cambridge: Polity Press.

Colley, H. & Jarvis, J. (2007). Formality and informality in the summative assessment of motor vehicle apprentices: a case study. *Assessment in Education*, **14**, 3, 295–314.

Cohen, L. & Mannion, L. (1994). *Research methods in education*. Fourth edition. London: Routledge.

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–961.

Greenhalgh, T. (1999). Narrative based medicine in an evidence based world. *British Medical Journal*, **318**, 323–325.

House, E. (1991). Realism in Research. *Educational Researcher*, **20**, 6, 2–9.

Johnson, M. (2007). Does the anticipation of a merit grade motivate vocational test takers? *Research in Post-Compulsory Education*, **12**, 2, 159–179.

Johnson, M. (*in press*). Exploring assessor consistency in a Health and Social Care qualification using a sociocultural perspective. *Journal of Vocational Education & Training*.

Kuhn, T. S. (1970).*The structure of scientific revolutions*. 2nd edition. Chicago: University of Chicago Press.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, **33**, 2, 3–11.

Oakley, A. (2000). *Experiments in knowing: gender and method in the social sciences*. Cambridge: Polity Press.

Oates, T. (2007). *The constraints on delivering public goods – a response to Randy Bennett's 'What does it mean to be a nonprofit educational measurement organization in the 21st Century?'* Paper presented at the IAEA conference, Baku.

Pope, C. & Mays, N. (1995). Qualitative research: reaching the parts other methods cannot reach. *British Medical Journal*, **311**, 42–45.

Rapport, F., Wainwright, P. & Elwyn, G. (2004). "Of the edgelands": broadening the scope of qualitative methodology. *Journal of Medical Ethics; Medical Humanities*, **31**, 37–42.

Schulenberg, J. L. (2006). Analysing police decision-making: assessing the application of a mixed-method/mixed-model research design. *International Journal of Social Research Methodology*, **10**, 2, 99–119.

Do assessors pay attention to appropriate features of student work when making assessment judgements?

Victoria Crisp Research Division

Introduction

This article draws on a study of the cognitive and socially-influenced processes involved in marking (Crisp, 2007; Crisp, *in press*; Crisp, *in submission*) and grading (analysis ongoing) A-level geography examinations and pilot research into the marking of GCSE coursework by teachers. These data were used to investigate the features of student work that examiners and teachers pay attention to and whether these features are always appropriate.

Where assessments involve constructed responses, essays or extended projects, the human judgement processes involved in assessing work are central to achieving reliable and valid assessment. Consequently, we need to know that appropriate features of student work influence assessment decisions and that irrelevant features do not.

Lumley (2002) suggests that less typical responses that are not accommodated in the assessment guidance force assessors to develop their own judgement strategies and they may be influenced by their intuitive impressions. If this is the case, there is the potential for criteria that are not intended to be used in marking to have an influence.

Several studies (Milanovic, Saville and Shuhong, 1996; Vaughan, 1991) have investigated marking processes in the context of English as a second language and key criteria used during assessment could be identified. Vaughan also found that different assessors (making holistic ratings) focus on different aspects of essays to each other and may have individual approaches to reading essays. Elander and Hardman (2002), in the context of psychology examinations, found that different examiners valued different factors more or less and that different factors were more predictive of the overall mark with different markers.

In the context of grading (or awarding) decisions, Cresswell (1997) found little evidence in awarders' verbalisations in meetings of how particular features of candidate work influenced decisions. Work by Murphy *et al*. (1995) found that awarders' individual views of what constitutes grade worthiness were more important in determining their decision making than other information such as statistics (although other information played a part). Further to this, Scharaschkin and Baird (2000) found that the degree of consistency of student work within a script, a feature that was not a part of the mark scheme guidance, influenced grading decisions for biology and sociology A-level scripts.

Sanderson (2001) developed a model of the process of marking A-level essays which emphasised (amongst other things) the social context of assessment judgements. Cresswell (1997) identified affective reactions to scripts (e.g. like or dislike) by examiners in awarding meetings. It is hypothesised that social, personal and affective reactions could perhaps affect the features attended to by assessors and explain some differences between examiners in terms of marks awarded.

The main focus of the research studies drawn on here was to improve our understanding of the judgement processes involved in marking and grading by examiners and marking by teachers. However, the focus of the additional analyses for this paper was on investigating whether assessors pay attention to appropriate features of student work when making assessment judgements.

Method

This article draws on data from two research studies both using verbal protocol analysis methodology. Verbal protocol analysis involves asking participants to complete a task whilst 'thinking aloud' and then using the verbalisations to infer the processes going on. This is generally considered a suitable method for investigating cognitive processes but has limitations in that certain types of information or processes do not occur at a conscious level and so can not be reported by participants (Ericsson and Simon, 1993).

The first set of data drawn on in this paper was collected in the context of A-level geography examinations and the main analyses have been reported in Crisp (2007; *in press*; *in submission*). Six experienced examiners were involved in the research and after some initial marking each examiner marked four to six scripts from each exam whilst thinking aloud. Each examiner also carried out a grading exercise for each exam whilst thinking aloud in which they were asked to judge the A/B boundary for the paper (i.e. to judge the minimum mark worthy of an A grade). During the grading exercise examiners had access to relevant parts of the Principal Examiner's report to the awarding team and had two scripts on each of the marks within the range used in the original awarding meeting. The grading exercises aimed to simulate and gain insight into the cognitive aspects of grading judgements without interference from the potential influence of social or political dynamics of live awarding meetings.

The second set of data drawn on in this paper was collected for pilot research in the context of GCSE coursework. One English teacher and one Information and Communications Technology (ICT) teacher each marked two coursework pieces at home and then later marked two further pieces whilst thinking aloud.

With both these sets of data the verbal protocols were analysed in detail using appropriate coding schemes (see, for example, Crisp, *in press*). A range of types of assessor behaviours and reactions were identified including reading behaviours, evaluations and personal, affective and social reactions.

With the A-level data the frequencies of different types of behaviours were compared between the exams and between examiners (see Crisp, 2007; Crisp, *in press*). Tentative models of the marking process and the grading process were developed by investigating patterns of behaviours/codes and the likely cognitive processes were considered in relation to existing theories of judgement (Crisp, *in submission*). This work

identified that evaluations either occurred alongside reading ('concurrent evaluations') and involved an evaluation of a part of the work, or occurred at a more overall level ('overall evaluations') and involved bringing together the understanding of the student's response, including its strengths and weaknesses, and beginning to convert this to a mark or grade decision (Crisp, *in submission*).

With the data from GCSE coursework marking, the teacher behaviours and reactions were compared between subjects (though with some caution given that there was only one teacher in each subject in this pilot work).

Results

For this article, additional analyses of the data were conducted. This involved reviewing extracts of the verbal protocol transcripts where assessors paid attention to particular features of student work or showed particular reactions, and then ascertaining whether these features affected evaluations. Evaluations were found to occur either concurrently with reading (usually an evaluation of a particular element of the student work) or after reading is complete as part of an overall evaluation and consideration of the appropriate mark. This distinction will be used to structure the analysis. This article focuses mostly on the data from A-level geography marking. It will consider data from the A-level geography grading exercises and the GCSE coursework marking pilot research more briefly.

Geography A-level marking and grading

Most aspects noted by examiners were closely related to the mark scheme and were about geography content knowledge, understanding and skills. Additionally, examiners sometimes made comments relating to aspects of students' attempts to achieve the requirements of the task ('task realisation') (see Crisp, *in press*). These included comments on the length of a response, noting whether the student had understood the question, commenting on the relevance of points and on material missing from a student's response (Crisp, 2007; Crisp, *in press*). Most of the features noted by examiners in this category are likely to be legitimate influences on examiner judgements. One exception might be the length of responses which probably should not affect marks directly. A further more detailed look at the verbalisations coded in this category revealed that all evaluative comments on length related to the response being shorter than expected and hence not showing sufficient knowledge, understanding and skills, or being longer than expected and including too much information that is not necessarily used to directly answer the question. In both cases it then becomes acceptable for these factors to affect examiner judgements as they are aligned with the marking criteria.

References to the geography A-level Assessment Objectives during marking were coded in the analysis (Crisp, 2007; Crisp, *in press*) as this gives insight into how examiners convert what they have seen (possibly categorising and combining cues or information) into marks. The high frequency of reference to Assessment Objectives (6.88 references to an Assessment Objective per script on average during marking) and the fairly frequent association with positive or negative evaluations (5.97 instances on average per script of a reference to an Assessment Objective co-occurring with a positive or negative evaluation) gives a strong indication that markers do tie their thinking closely to the valued aspects of the mark scheme guidance (i.e. the intended marking criteria). There was also fairly frequent reference to the mark scheme during marking (2.03 times on average per script). The analysis will now focus on aspects of marker verbalisations that were less expected and less clearly related to the qualities described in the mark scheme.

Language

Examiners sometimes commented on the quality of a student's language use or on orthography (i.e. handwriting, legibility and presentation) (see Crisp, 2007; Crisp, *in press*). This occurred 1.46 times per script on average during marking. A more detailed analysis of the marking transcripts for each of the 86 instances revealed that 27 instances were not associated with any evaluation, 58 instances were associated with either a positive or negative concurrent evaluation (i.e. an immediate evaluation made during the process of reading the response), 24 instances fed into overall evaluations relating to Communication as an Assessment Objective, and 10 instances were associated with overall evaluations that were not specifically linked to assigning marks for communication[1].

This suggests that language quality rarely impacts on overall evaluations except where communication is an explicit criterion for evaluation (as in the A2 exam). Instances where reference to language use did feed into overall evaluations occurred where the structure was weak resulting in a reduced clarity in the student's meaning or where the legibility of the response was sufficiently weak to impair understanding of the student's meaning and line of argument. It seems that language only affects overall evaluations where communication is an aspect intended to be assessed or in circumstances where the quality of language or handwriting impairs understanding.

It is interesting that in a number of the instances where language quality or orthography was associated with a concurrent evaluation examiners said that a response would get a certain number of marks *despite* its weak structure or expression. This might suggest that they are in control of the influences on their marking and prevent language skills from impacting their judgements where marking guidance determines that it should not.

Of the 28 instances of reference to language use during grading, 22 were associated with a concurrent evaluation (e.g. 'sound introduction, quite well written') and 7 were associated with the overall evaluation of the quality of the script. In the instances that fed into overall evaluations it seems that language quality was occasionally one factor in the examiner's mind when attempting to make a judgement of grade worthiness even when it was not an explicit mark scheme criterion. However, it is interesting to note that all comments on language which seemed to feed into overall evaluations were positive rather than negative.

Social perceptions

As noted in Crisp (Crisp, *in press*) examiners sometimes appear to have social perceptions of students during marking as understood from characteristics of the script. Markers sometimes made assumptions about other characteristics of students (0.85 per script on average) or inferred likely further performance of the student (0.39 per script on average).

The code 'assumptions about candidates' was applied where an examiner inferred student characteristics (e.g. ability, lazy, thoughtful)

1 In this and the analyses that follow some instances of a particular code were associated with both a concurrent and an overall evaluation. Consequently the numbers quoted sometimes add up to more than the total number of instances.

or inferred how a student has approached the task from the student's response. Reviewing transcript extracts revealed that assumptions about candidates were often about general geography ability or specific aspects of knowledge (e.g. knowledge of place) and were hence part of the examiner's progress towards forming an overall impression of a student's relevant abilities. Detailed analysis of the 50 instances of this code found that 17 instances were not associated with an evaluation, 26 instances were associated with a positive or negative concurrent evaluation, and 26 instances were issues that fed into overall evaluations and so may have influenced the marks awarded. Of the 26 instances of assumptions about candidates being linked to overall evaluations 23 were at least partly about the student's geography ability or knowledge, for example: '*this lad knows a lot, likes to write a lot*'. The three instances linked to overall evaluations that did not relate to geography ability still related closely to the students' attempts to answer the questions.

In grading, assumptions about candidates were infrequent (0.13 times per script on average or 12 instances in total). In a similar way to during marking, instances sometimes related to concurrent evaluations (5 instances) or overall evaluations (3 instances) but were usually assumptions relating to geography abilities or to do with the students' attempts to answer the questions. As with marking, such assumptions seem to aid the examiner in synthesising their understanding of different aspects of the student's response in order to come to an understanding of the overall level of performance.

Examiners occasionally made predictions about candidate performance before finishing reading a response or sometimes even before beginning to read (Crisp, 2007; Crisp, *in press*). Predictions related to the likely quality of the response or to the kinds of material they expected to see in the rest of the response or script, for example: '*This is not going to be a better paper, is it?*'

Analysis of the 23 instances of performance predictions (from the marking protocols) found that 7 involved no evaluation, 16 included a concurrent evaluation (e.g. '*not going to be a strong script I think*') and 5 were associated with considering the overall performance. Where predictions are associated with the overall evaluations these often occurred later in the reading of a response (when the examiner has more information and so it is more reasonable for them to make an overall prediction). The rest of the response was still read carefully and the entire view of the script was checked against the marking criteria.

There were very few instances of examiners predicting performance in the grading data (0.04 per script on average) and these were similar in nature to the instances during marking (expecting certain content, hoping response will get better). Only 1 of the 4 instances contained an evaluation in grading and this was a concurrent rather than an overall evaluation.

Personal and affective reactions

Examiners sometimes showed affective (i.e. emotional) or personal reactions to features of students' work (Crisp, 2007; Crisp, *in press*). During marking, positive affect (e.g. '*so good he is on target now, I'm really pleased*') was shown 0.75 times per script on average and negative affect was displayed 1.24 times per script on average. Examiners showed amusement or laughed during marking 0.49 times per script on average and showed frustration 0.39 times per script on average.

There were a total of 44 instances in total of examiners showing positive affect (or sympathy) towards students and/or their work during marking. Of these, 20 instances were not associated with an evaluation, 20 were linked to a concurrent evaluation and 5 were linked to an overall evaluation. Instances of positive affect being linked to concurrent evaluations usually involved a positive feature of a script eliciting both a positive evaluation and positive affect (e.g. '*oh hooray, hooray, hooray, someone has actually thought about that!*') or a feature of the script eliciting sympathetic feelings and a negative evaluation. In both types of instances it is the positive or negative evaluation and not the examiner's affective reaction which may be going on to influence further evaluation.

In grading, evidence of positive affect was fairly infrequent and the verbalisations showing positive affect were similar in nature to those occurring during marking.

There were 73 instances of examiners showing a negative affective reaction to student work (e.g. '*oh no not the flippin' Italian dam again*') during marking. Of the instances, 41 were not associated with any evaluation, 27 were associated with a concurrent evaluation and 6 were associated with an overall evaluation. Looking at the instances of links with concurrent and overall evaluations suggests that, similarly to positive affect, negative affect is usually a response to negative aspects of students' responses in terms of the knowledge and skills required, or a response to efforts to appropriately answer questions. Some verbalisations also indicated that examiners were sufficiently aware of their emotional responses to not allow these to influence the marks they award. Negative affective reactions were infrequent in grading. Most instances were not associated with evaluations and those that were, were similar in nature to the instances in marking.

In marking, there were 29 instances of laughter or amusement in response to student work. Only 6 instances were linked to concurrent evaluations and none to overall evaluations. The concurrent evaluations tended to occur where a student gave certain kinds of factually incorrect information which are then evaluated as incorrect. Amusement and laughter were infrequent in grading and were only associated with a concurrent evaluation on one occasion.

Frustration or disappointment was shown by examiners in 23 instances in relation to marking. In 7 instances this was not connected to evaluations, in 13 it was linked to a concurrent evaluation and in 4 instances to an overall evaluation. Where examiners showed frustration or disappointment linked to a concurrent or overall evaluation this tended to be where the student's work was weak in some respect, something was missing from their response or their response was not appropriately targeted to the question. In grading frustration was infrequent. As with marking more than half of these instances were related to some kind of evaluation but they appeared to relate to legitimate weaknesses in student work.

It seems that although a number of different types of emotive reactions were elicited from examiners, these affective responses were caused by qualities of the geography or students' abilities to achieve the task, and it was this rather than any emotional response that guided marking and grading decisions.

GCSE coursework marking

This section will describe briefly the features attended to by teachers when marking GCSE coursework using the pilot study. These data do need to be treated with some caution due to the small scale of this pilot work but may provide insight into whether the findings in A-level geography are likely to generalise to marking by teachers, marking in other subject areas and marking of a different type of student work.

First, it is worth noting that the teachers referred to the marking guidance fairly frequently, and particularly frequently in ICT (19.5 times per coursework piece for ICT and 3.5 times per coursework folder in English on average). The difference in frequency between subjects relates to the nature of the mark schemes. The ICT mark scheme includes very specific task elements that students need to show in their work, and hence requires very close reference to the mark scheme during marking. The mark scheme for the English coursework represented a continuum on a number of different types of skills and thus appears to be easier for teachers to internalise, such that they do not need to refer to it as frequently.

In the pilot work it was considered useful to code the detailed features of student work commented on by teachers in their verbalisations to allow investigation of differences between subjects. In English these included:

- evaluates spelling, punctuation or grammar
- evaluates style, vocabulary, quality of expression, use of technical terminology or text structure
- evaluates imagination, sophistication, whether interesting or formulaic
- student's personal response to literary texts
- making comparative points about texts/poems
- understanding of genre
- student's use of quotations from literature
- presence of/quality of conclusions to essays
- use of narrative

In ICT features focussed on included:

- evaluates spelling, punctuation or grammar
- evaluates style, vocabulary, quality of expression, use of technical terminology or text structure
- use of IT and non-IT source materials
- absence/presence of information or evidence on the sources used
- designs/image editing
- saving files and folders
- use of number
- spell-checking and proof-reading

These are all features included in the relevant marking criteria and are hence intended and legitimate influences on marking decisions.

Again there were other behaviours (either features of the work being noted or reactions occurring in response to features of the work) apparent in the transcripts which are less obviously related to intended influences on marking. These were similar to those seen in A-level exam marking and included:

- commenting on orthography
- commenting on aspects of task realisation (e.g. response length)
- affective reactions and amusement
- social perceptions (e.g. predicting performance, reflections on characteristics of students)

Looking at the verbalisations fitting these codes suggests that, similarly to the marking and grading of A-level geography, inappropriate features of student work do not appear to influence evaluations in ways that they should not.

Discussion

The verbal protocol methodology was generally a successful method for exploring the features of student work attended to during marking. However, the limitation of the method in terms of verbal protocols not supplying a complete record of all thoughts passing through working memory (Ericsson and Simon, 1993) is problematic. Therefore, we cannot be completely sure that no inappropriate features of student work ever influenced overall evaluations and mark decisions in unintentional ways although the data are encouraging in this respect.

The data collected suggest that assessors mostly attend to features of student work related to intended marking criteria during their marking or grading process and that they focus mostly on the intended marking criteria in their actual evaluations. Most of the verbalisations focussed on features relevant to the subject knowledge, understanding or skills under assessment and Assessment Objectives and the marking guidance were used fairly frequently. There were, however, some types of behaviours or reactions during their processing that might, at first inspection, indicate that assessors sometimes attend to features of student work that are not within the intended focus of evaluations. Analysis of these instances revealed that where features were attended to that were not indicated by the mark scheme these did sometimes influence ongoing evaluations and occasionally fed into overall evaluation and mark consideration. However, close analysis indicated that most instances were actually caused by features of the student work that were intended to be evaluated. Additionally, several verbalisations indicated that although features were noted and sometimes considered during evaluations, assessors tended to be in control of whether these influenced actual marks.

Given that inappropriate features of student work and personal, social and affective reactions did not appear to influence overall evaluations and mark consideration inappropriately, it seems that such behaviours do not explain variations in marks between examiners. This may suggest that variations are a result of other factors perhaps such as variations in the weight that examiners place on different features, variations in the extent to which examiners are willing to be lenient when inferring a student's knowledge behind a partially ambiguous response, or variations in the interpretation of aspects of the mark scheme. These issues would require further investigation to ascertain their contribution.

The data are consistent with the view that the judgement processes involved in the assessments investigated rely closely on professional knowledge and that evaluations of work are strongly tied to values communicated by the mark scheme. Features relating to task realisation also legitimately influence evaluations. Thoughts regarding language use, social perceptions and affective reactions also sometimes led to concurrent evaluations and occasionally fed into overall evaluations but assessors were in control of influences on their judgements and no inappropriate biases were found using the current methods.

Note:

This article is based on a paper presented at the International Association for Educational Assessment Annual Conference in Baku, Azerbaijan, September 2007.

References

Cresswell, M. J. (1997). *Examining judgements: theory and practice of awarding public examination grades*. PhD Thesis. Unpublished doctoral dissertation, University of London, Institute of Education, London.

Crisp, V. (2007). *Comparing the decision-making processes involved in marking between examiners and between different types of examination questions*. Paper presented at the British Educational Research Association Annual Conference, London.

Crisp, V. (*in press*). Exploring the nature of examiner thinking during the process of examination marking, *Cambridge Journal of Education*.

Crisp, V. (*in submission*). Towards a model of the judgement processes involved in examination marking.

Elander, J. & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology*, **93**, 303–328.

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, **19**, 246–276.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behaviour of composition-markers. In: M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE awarding*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

Sanderson, P. J. (2001). *Language and differentiation in examining at A Level*. PhD Thesis. Unpublished doctoral dissertation, University of Leeds, Leeds.

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

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In: L.Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J: Ablex Publishing Corporation.

Marking essays on screen: towards an understanding of examiner assessment behaviour

Stuart Shaw CIE Research

Introduction

Computer assisted assessment offers many benefits over traditional paper methods. In translating from one medium to another, however, it is crucial to ascertain the extent to which the new medium may alter the nature of the assessment and marking reliability. Appropriate validation studies must be conducted before a new approach can be implemented in high stakes contexts. The pilot described here is the first attempt by Cambridge International Examinations (CIE) to mark, on-screen, extended stretches of written text for the Cambridge Checkpoint English Examination. The pilot attempts to investigate marker reliability, construct validity and whether factors such as annotation and navigation differentially influence marker performance across the on-paper and on-screen marking modes.

Candidates wrote their answers on paper scripts in the normal way. The scripts were then scanned and digital images of them were sent by secure electronic link to examiners for on-screen marking at home using Scoris® software.

It can be relatively hard for examiners to make a full range of annotations when marking on screen. For this reason annotation sophistication was manipulated in the pilot as well as marking mode. Four marking methods were compared: on-paper with sophisticated annotations (current practice), on-paper with simplified annotations, on-screen with sophisticated annotations, and on-screen with simplified annotations.

The research literature

There is a large research literature relevant to this project. Key aspects of this literature are summarised below.

Comparability of marking across on-screen and on-paper modes

The literature is mixed on this topic.

- Bennett (2003) carried out an extensive review of the literature and concluded that 'the available research suggests little, if any, effect for computer versus paper display' (p.15).
- Differences were found in a few studies not reviewed by Bennett, however, e.g. Whetton and Newton (2002) and Royal-Dawson (2003).
- Sturman and Kispal (2003) observed quantitative differences between online and conventional marking of tests of reading, writing and spelling for pupils typically aged 7 to 10 years, but an analysis of mean scores showed no consistent trend in scripts receiving lower or higher scores in the e-marking or paper marking: 'absence of a trend suggests simply that different issues of marker judgement arise in particular aspects of e-marking and conventional marking, but that this will not advantage or disadvantage pupils in a consistent way' (p.17). Sturman and Kispal concluded that e-marking is at least as accurate as conventional marking. Wherever differences between the

two marking modes existed they tended to occur when marker judgement demands were high. They also noted that when assessing a pupil's response on paper, holistic appreciation of the entire performance may contribute to a marker's award, but this is not possible if scripts are split up by question for on-screen marking.

- Shaw, Levey and Fenn (2001) have investigated the effects of marking extended writing responses across modes. Scripts from Cambridge ESOL's December 2000 Certificate in Advanced English examination, were scanned and double-marked on-screen. Statistical analysis of the marking indicated that examiners awarded marginally higher marks on-screen and over a slightly narrower range of scores than on paper. The difference in marking medium, however, did not appear to have a significant impact on marks.
- Twing, Nichols, and Harrison (2003) also looked at extended prose on screen. The allocation of markers to groups was controlled to be equivalent across the experimental conditions of paper and electronic marking. Findings revealed that marks from the paper-based system were slightly more reliable than from the screen-based marking. The researchers canvassed opinion from markers and deduced that for some, interaction with computers was a new experience. For these markers, lack of computer experience and familiarity engendered anxiety about on-screen marking. Research suggests that anxiety over computer use could be an important factor militating against statistical equivalence (McDonald, 2002). Mere quantity of exposure to computers is not sufficient to decrease anxiety (McDonald, citing Smith, Caputi, Crittenden, Jayasuriya and Rawstorne 1999) – it is important that users have a high quality of exposure also. Interestingly, for those markers experienced with computers, Twing *et al*. (2003) found that image-based markers finished faster than paper-based markers.
- The question of whether examiners make *qualitatively* different judgements when marking the same piece of writing in different marking modes is a key consideration in assessment (Shaw and Weir, 2007). There is very little research to draw upon in this area. Johnson and Greatorex (2006) conclude that judgements made on-screen and conventionally on paper are qualitatively different, stressing that effects of mode on assessment evaluations are both important and in need of on-going inquiry.
- Although much evidence suggests that examiners' on-screen marking of short answer scripts is reliable and comparable to their marking of the paper originals, it is clear that more research is needed, particularly concerning assessment of extended responses on-screen, to ascertain in exactly what circumstances on-screen marking is both valid and reliable.

Examiners' annotations

- There is a relative paucity of literature relating to the use, purpose and application of annotations in examination marking.
- Crisp and Johnson (2005) suggest that annotations serve two distinct functions: as an accountability function (*justificatory*) and as a means of supporting examiners' decision-marking processes (*facilitation*).

Justificatory function

- Murphy (1979) notes that senior examiners are influenced by the marks and comments on scripts during the process of review marking.
- In their experimental study on the use of annotations in Key Stage 3 English marking, Bramley and Pollitt (1996) observed that 'having annotations on the scripts might enable team leaders to identify markers whose marks need checking' (p.18).
- As part of an investigation into marking reliability involving double marking, Newton (1996) explored whether correlations between first and second marks were affected by obscuring the first marker's comments from the second marker. Newton presented second markers with 'partially obscured' scripts, where the first marker's marks had been obscured but the comments left visible, and 'fully obscured' scripts, where both marks and comments had been obscured. The correlation between first and second marks was a little higher for the partially obscured scripts, but the difference did not reach statistical significance.
- Williamson (2003) asserts that annotations might have an important communicative role in the quality control process.

Facilitation function

- Bramley and Pollitt (1996) observed that the majority of markers considered that annotating contributed to the improvement of their marking, helped them to apply performance criteria, and reduced the subjectivity of their judgements.
- O'Hara and Sellen (1997) suggest that readers of texts annotate in order to highlight structural features of the text and salient features, to record questions or draw attention to ideas that require reflection or further investigation.
- Annotations may offer cognitive support for comprehension building as well as performing other functions which are specifically linked to the context of the examination process (Anderson and Armbruster, 1982; Askwall, 1985; O'Hara, 1996; O'Hara and Sellen, 1997; Benson, 2001; Crisp and Johnson, 2005);
- According to Bramley and Pollitt (1996, p.6), 'Annotating might reduce the cognitive load of markers during the judging process by creating a "visual map" of the quality of an answer, assisting comparisons with other answers'.
- In assessing feedback given to students when assignments were submitted and feedback returned on paper as well as on screen, Price and Petre (1997) observed that the quality and type of feedback were found to be similar. However, annotations providing *emphasis* were used less on-screen (although their use increased with increasing software familiarity).
- Shaw (2005) observed that examiners use annotations to investigate their own marking consistency. Annotations provide an efficient means to confirm, deny or reconsider standards both within and across candidates thereby reassuring examiners throughout the marking event.
- Crisp and Johnson (2005) investigated the use of annotations made by examiners marking a small number of GCSE Mathematics and Business Studies scripts. Their findings indicated that markers consider annotating to be a positive aspect of marking. This reflects the conclusions drawn by Bramley and Pollitt (1996) which suggest

that markers understand the process of annotations as being integral to, and contributing towards, the efficacy of marking.

Reading on-screen

- A growing body of research suggests that reading strategies employed to achieve comprehension of essays on paper play a vital role in the marking process and hence have implications for the reliability of marking (Sanderson, 2001; Crisp, 2007; Suto and Nádas, *in press*).
- Reading on-screen is 'generally less appealing than reading from paper' (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt and Schedl, 2000, p.41).
- Research on first language (L1) reading indicates that reading rates drop 10–30% when moving from printed material to on-screen reading (Muter and Maurutto, 1991; Kurniawan and Zaphiris, 2001). Segalowitz, Poulsen and Komoda found that second language (L2) reading rates of highly bilingual readers are '30% or more slower than L1 reading rates' (1991, p.15).
- No single factor can account for why reading on-screen is perceived to be more difficult than reading on paper. In fact a number of variables are associated with reading on-screen: screen resolution, spatial representation, ease of use, disorientation, non-tangibility, experience, etc.
- Cassie (undated) cites two reasons why reading may be more difficult on a computer screen than on paper. First, readers tend to relate certain topics with strategically-situated locations on the page where they appear. Secondly, the process of reading through a number of printed pages is a tactile one: the reader having some comprehension of how far they have 'travelled' through a document.
- Related research has investigated the effects of computer familiarity on on-screen reading (Kirsch *et al*, 1998) and the effects of screen layout and navigation on reading from screen (Dyson and Kipping, 1998; dos Santos Lonsdale, Dyson and Reynolds, 2006).
- The visual layout of text and the mode of presentation affects the ease with which readers can access, read and respond to the text (Foltz, 1993; O'Hara and Sellen, 1997).
- Prior reading experience and computer familiarity are among factors that can influence reading assessment and methods (Rothkopf, 1978; Rayner and Pollatsek, 1989).
- Most empirical research into reading on-screen has separately addressed manipulation or navigation e.g. document structure, scrolling, page management (McDonald and Stevenson, 1996; Wenger and Payne, 1996; McDonald and Stevenson, 1998a, 1998b; Lin, 2003) and visual ergonomic factors e.g. layout variables (Dillon, 1994, 2004).
- One element of scrolling patterns (pauses between scrolling movements) has been identified as the main determinant of reading rate on-screen (Dyson and Haselgrove, 2000).

Context of the pilot

The Cambridge Checkpoint English examination is an innovative diagnostic testing service which provides standardised assessments for mid-secondary school pupils aged around 14. The tests, offered at two sessions each year, are designed to give feedback on individual strengths and weaknesses in the key curriculum areas of English, Mathematics and Science. The results provide teachers with information on student performance, enhanced by reporting tools built into the Checkpoint service.

English is assessed using two papers. Each paper takes one hour with an additional seven minutes for reading. In terms of the writing requirements, in Paper 1 candidates are given a short, focussed task with a clear aim and audience. The content is non-narrative and candidates are expected to write about 250 words. Paper 2 consists of a short and focussed task that does have a narrative content. Again, candidates are expected to write about 250 words.

Pilot design

The pilot employed a mixture of quantitative and qualitative methods. Quantitative methods used included correlational analyses of marks; computation of examiner inter-rater reliabilities; and Multi-Faceted Rasch Analyses (MFRA). The qualitative dimension of the pilot involved collating and analysing retrospective data captured by an examiner questionnaire. The research design, which was 'matched, between groups', tested the effect of two variables: marking medium and annotation sophistication, using four discrete marking conditions:

a) pilot scripts, **paper** marked, using **sophisticated** annotation
b) pilot scripts, **paper** marked, using **simplified** annotation
c) pilot scripts, marked **on-screen**, emulating current **sophisticated** annotation
d) pilot scripts, marked **on-screen**, using **simplified** annotation.

Table 1: Research Design

	Marking medium (Variable 1)		*Annotation (Variable 2)*	
	Paper	*On-screen*	*Sophisticated*	*Simple*
Method A	✔		✔	
Method B	✔			✔
Method C		✔	✔	
Method D		✔		✔

Ten examiners, including the Principal Examiner (PE), took part in the study, which consisted of two phases of marking. In phase 1, the examiners all marked the same set of 20 scripts on paper using sophisticated annotations. This 'calibration marking' provided a common baseline for the variation between these examiners under normal marking conditions. In phase 2, the examiners were split into four different sub-sets, one for each of the four marking conditions. All examiners then marked a further 200 scripts. Once again, the examiners marked the same scripts as each other (See Figure 1).

The examiners had various levels of experience but all had marked these question papers in the May 2007 administration and had been standardised then. The research was conducted in September 2007.

Marks and annotations from the live, on-paper May 2007 marking were removed from the 20 scripts which were subsequently coded,

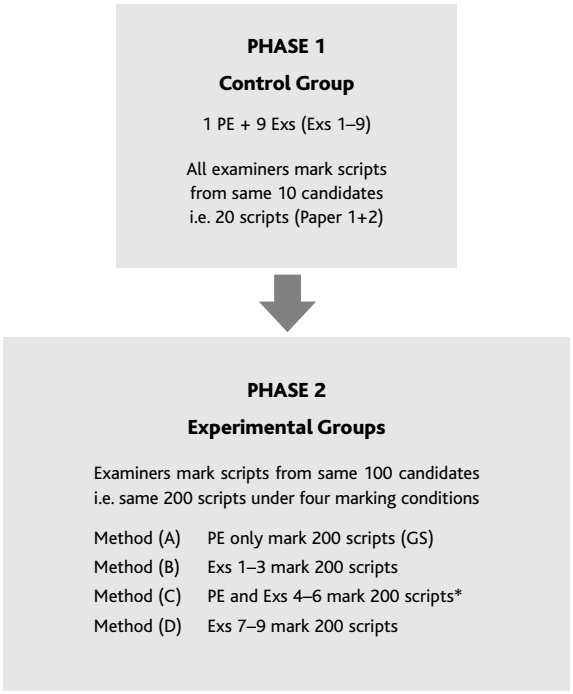


Figure 1: Research Design

copied and despatched to examiners for phase 1 of the pilot. The number of scripts required for the second phase of marking was arrived at through power test considerations (Kraemer and Thieman, 1987). Two hundred scripts (100 candidate performances) were scanned without annotations or marks to meet the requirements of marking under conditions described by Methods (C) and (D). In addition, unmarked hard copy versions were produced for Methods (A) and (B). Writing performances were identified as scripts which represented the full proficiency continuum for the test, exemplified a range of 'marked' profiles, and a diversity of centres.

In addition to empirical methodologies, emphasis was also attached to qualitative approaches. It was hoped that feedback from examiners would provide valuable insight into their on-screen marking experiences.

Findings

Phase 1: calibration markings

Descriptive statistics and analysis-of-variance indicated that the examiners were generally homogeneous in the marks they awarded to the 20 phase 1 scripts. Examiner inter-correlations were consistently high and indicated that examiners were reliably distinguishing between the respective assessment criteria on each paper. Strength of agreement tests revealed that whilst examiners were in general agreement on the rank ordering of the scripts, they were in less agreement regarding the absolute mark assigned to those scripts. However, inter-rater reliabilities were consistently high (of the order of 0.8), and Multi-Facet Rasch Analysis revealed that all examiners fell within the limits of acceptable model fit and that differences in severity / leniency between examiners were within tolerance (recommended cut off for flagging misfits includes t values outside +/- 2.0 [Smith, 1992]). The results of the phase 1 calibration markings therefore provide evidence that any quantitative differences found between the sub-groups in phase 2 are unlikely to be due to inherent differences between the markers in the sub-groups.

Phase 2: the four experimental marking methods

Before the marks from the four sub-groups were compared with each other, a quick comparison was made between the phase 1 and phase 2 marks. This indicated that examiners retained their relative levels of severity/leniency across both phases, that is, an examiner who was a little severe or lenient compared to the Principle Examiner in phase 1 was also a little severe or lenient in phase 2. As previously noted, however, there were no large differences in severity or leniency between examiners in phase 1.

Table 2 shows descriptive statistics across all four marking methods and for the live marks awarded in May 2007. The pilot means tended to be slightly higher than the live means.

The pilot standard deviations tended to be a little smaller than the live standard deviation for paper 1, but a little larger for paper 2. There were no large differences, however.

Table 3 shows the distribution of differences between the Principle Examiner marks for Method A (conventional marking) and the other examiners, aggregated by marking method. Method C (on-screen, sophisticated annotations) demonstrates the highest proportion of marks within +/- 3 marks of the PE.

Inter-examiner reliability indices were computed following the approach advocated by Hatch and Lazaraton (1991). A Pearson correlation matrix was generated for each marking method and then the average correlation for each method was calculated. A Fisher Z transformation was applied to the correlations before averaging to transform the correlations to a normal distribution suitable for averaging (Hatch and Lazaraton 1991). Table 4 presents the average correlations. The figures are high for both on-paper marking (method B) and on-screen marking (methods C and D). Although the inter-rater reliability is a little lower for the on-screen marking methods, the difference is not statistically significant.

Table 2: Overall comparison between Methods A – D and the live marks (Descriptive Statistics)

	Live May 2007			*Method A*			*Method B*			*Method C*			*Method D*		
	P1	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*	*P1*	*P2*	*Tot*
Mean	16.91	15.94	32.85	17.16	17.16	34.32	16.79	16.32	33.11	17.18	15.90	33.08	17.89	17.03	34.92
Std. dev.	6.71	6.00	12.10	6.12	6.14	11.69	6.54	5.96	11.49	6.28	6.20	11.81	5.57	5.94	10.70

Table 3: Agreement levels between the PE and other examiners

Marking Method	*Percentage of scripts:*			
	Exact agreement	*Within +/- 1 mark of PE*	*Within +/- 2 marks of PE*	*Within +/- 3 marks of PE*
Method B				
Paper 1	17	48	68	81
Paper 2	14	31	50	72
Method C				
Paper 1	21	52	71	82
Paper 2	13	32	47	80
Method D				
Paper 1	11	31	54	70
Paper 2	9	33	55	73

Table 4: Inter-examiner reliabilities

	Average correlation between examiners		
	Method B	*Method C*	*Method D*
Paper 1	0.80	0.78	0.75
Paper 2	0.80	0.78	0.78
Total (Paper 1 + Paper 2)	0.81	0.79	0.79

Findings from the retrospective questionnaire given to participants indicated that:

- Reading on-screen imposes higher cognitive demands on the marking process, particularly in relation to scrolling, page management, and application of annotations. Examiners suggested that protracted script electronic accessing procedures and slow script downloads may have deleterious consequences for the marking process. Pilot participants noted that their marking productivity was dependant upon several factors but chiefly the script downloading time.
- Examiners found scripts on-screen to be less easy to read than their paper counterparts (although this was not too great a problem for Checkpoint responses).
- Reading on-screen may adversely affect examiner concentration. Not being able to replicate paper and pen practice when applying annotations was a concern amongst pilot examiners. It was generally felt that on-screen marking is physically more demanding than paper marking and that marking over prolonged periods would engender mental and physical fatigue. For example, the physical process of selecting and applying pre-set annotations had implications for examiner concentration. It was believed that the additional cognitive demand intrudes upon the assessment process.
- Navigational demands imposed on the examiner by the computer interface affect the reading of text on-screen. Scrolling, for example, was considered by many examiners to be slow and generally annoying, presenting an unnecessary distraction to the reader.
- Script navigation was not as easy electronically as it is on paper. Reading on-screen inhibits formulation of a sense of overall meaning from the text and appears to impact negatively on examiner understanding of the marking criteria. Assessment criteria most affected tend to be those that define the macro features of text such as *rhetoric* (relating to discoursal features) and *organisation* (relating to coherence and cohesion).
- Whole text appreciation is impaired on-screen due to limited screen view and disrupted spatial layout. Holistic appreciation of the text was less achievable electronically as snapshots allow only restricted and incomplete sight of the text. This was especially noticeable when examiners were asked to consider the overall clarity and fluency of the message and how the response organises and links information, ideas and language.
- Reading on-screen may interfere with conventional, paper-based strategies employed to facilitate comprehension of the text message. The effect of mode seemed to encourage the use of different reading strategies, examiners having to revise their approach to assessment when marking on-screen.
- Prior experience with on-screen marking seems to have a positive influence on reading comprehension. Two of the pilot examiners, both of whom were consistent and reliable in their assessments (on paper and on-screen), claimed previous familiarity with on-screen marking.
- Identifying key features of textual information on-screen is more difficult than on paper.
- Reading on-screen may impede examiner construction of a mental representation of the text.
- Annotations aid textual comprehension. Whilst annotations are more awkward to apply on-screen, examiners were universal in their assertion that inability to annotate may impact negatively on the marking process. Participants were unanimous in their belief that the process of annotating enabled them to arrive at the right judgement(s).
- On-screen annotating may enhance marker reliability particularly as the software imposes a standardised set of electronic annotations.
- Examiners using the simplified form of annotation did not consider the range of annotation sufficient for marking purposes: the simplified suite of annotations being too restrictive.
- Examiners reinforced the prevailing belief that annotated scripts serve as a permanent record for subsequent adjudication and perform a communicative function between examiners.
- Generally, examiners were mixed regarding whether the time taken

to mark scripts on screen was the same as the time required to mark ordinary paper scripts. Despite difficulties encountered both reading and assessing on-screen, the majority of examiners believed that they ended up with about the same mark for each candidate across both modes. Whilst most examiners would still prefer to mark on paper, finding on-screen marking less enjoyable, nearly all examiners would be willing to use similar software in future sessions.

Discussion and Conclusion

The pilot found that paper-based and screen-based inter-examiner reliability is high for the Cambridge Checkpoint English Examination. Although inter-rater reliability is lower on-screen it is only marginally deflated. This finding accords with the findings of other, similar studies (e.g. Twing *et al*., 2003).

Levels of agreement were investigated between the Principle Examiner, marking on paper using sophisticated annotations, and other examiners marking on paper with simplified annotations, on-screen with sophisticated annotations, and on-screen with simplified annotations. The best agreement was found for those examiners marking on-screen with sophisticated annotations, implying that using sophisticated annotations is more important for marking accuracy than whether the marking is done on screen or on paper.

Analysis of mark agreement can only take us so far in an investigation of comparability, however, since a high degree of mark convergence might still mask issues to do with construct validity. This might be because the scripts used in the study did not cover the full range of relevant features, or because the examiners were not marking correctly in either mode.

Construct validity refers to the extent to which the testing instrument measures the 'right' underlying psychological traits or 'constructs'. Clearly, it is important to ensure that the constructs that tests are measuring are precisely those they intend to and that these are not contaminated by other irrelevant constructs or effects. If the mode of marking or the level of annotation permitted affect examiners' reading or understanding of the text, their assessments may be affected and construct validity compromised.

A reasonably well-developed conceptualisation of construct validity encompasses three dimensions of any testing activity – cognitive validity (the cognitive processing by the candidates activated by the test question), context-based validity (consideration of the social and cultural contexts in which the question is performed as well as the content parameters) and scoring validity which relates to all aspects of reliability (Shaw and Weir, 2007). If aspects of scoring validity are compromised by different modes of presentation then construct validity is potentially threatened. The questionnaire data collected in the present study revealed a number of functional differences between on-screen and on-paper marking modes, and between simple and sophisticated annotations, that might affect construct validity, and these would repay further investigation.

Future research

Future research should aim to:

- Establish the effects of navigation facilities and annotative tools on reading assessment, particularly in the context of longer stretches of text.
- Identify conditions under which examiner assessment is affected by interface design.
- Develop a greater knowledge of reading processes on-screen through:
  - identifying means by which differences in reading are mediated;
  - exploring whether reading can be enhanced by manipulating mediating factors.

CIE will undertake future pilots with these aims in mind. Reliability across marking mode will continue to be an important consideration. One study will entail marking the Singapore General Paper (GCE AO Level) Paper 2 on-screen. Paper 2 includes two questions that, in terms of expected text length, make greater demands on candidate resources than the Checkpoint English test. In general, the longer the text candidates have to produce, the greater the language, content knowledge, organisational and monitoring metacognitive abilities that might be required in processing. Concomitant with these demands on candidates is an increased cognitive load placed upon the assessor during marking.

References

Anderson, T. H. & Armbruster, B. B. (1982). Reader and text-studying strategies. In: W. Otto & S. White (Eds.), *Reading Expository Material*. London: Academic Press.

American Psychological Association (1985). *Guidelines for Computer-based Tests and Interpretations*. Washington, DC: Bennett.

Askwall, S. (1985). Computer supported reading vs. reading text on paper: a comparison of two reading situations. *International Journal of Man Machine Studies*, **22**, 425–439.

Bennett, R. E. (2003). *On-line Assessment and the Comparability of Score Meaning* (ETS RM-03-05). Princeton, NJ: Educational Testing Service.

Benson, P. J. (2001). Paper is still with us. *The Journal of Electronic Publishing*, **7**, 2. Available online at: www.press.umich.edu/jep/07-02/benson0702.html (accessed 25 August 2005).

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Bramley, T. & Pollitt, A. (1996). *Key Stage 3 English: Annotations Study*. A report by the University of Cambridge Local Examinations Syndicate for the Qualifications and Curriculum Authority. London: QCA.

Cassie, T. (undated). *Reading and navigating of documents: digital versus paper*. Department of Computer Science, University of Maryland.

Crisp, V. (2007). Researching the judgement processes involved in A-level marking. *Research Matters: A Cambridge Assessment Publication*, **4**, 13–18.

Crisp, V. and Johnson, M. (2005). *The use of annotations in examination marking: opening a window into markers' minds*. Research Programmes Unit, Cambridge Assessment. A paper presented at the British Educational Research Association Annual Conference, University of Glamorgan, September 2005.

Dillon, A. (1994). *Designing usable electronic text: ergonomic aspects of human information usage*. London: Taylor and Francis.

Dillon, A. (2004). *Designing usable electronic text: ergonomic aspects of human information usage*. 2nd edition. London: CRC Press.

dos Santos Lonsdale, M., Dyson, M. C. & Reynolds, L. (2006). Reading in examination-type situations: the effects of text layout on performance. *Journal of Research in Reading*, **29**, 4, 433–453.

Dyson, M. C. & Haselgrove, M. (2000). The effects of reading speed and reading patterns on our understanding of text read from screen. *Journal of Research in Reading*, **23**, 1, 210–223.

Dyson, M. C. & Kipping, G. J. (1998). The effects of line length and method of movement on patterns of reading from screen. *Visible Language*, **32**, 2, 150–181.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series 17.

Foltz, P. W. (1993). *Readers' comprehension and strategies in linear text and hypertext*. Unpublished doctoral dissertation, University of Colorado, Boulder. Cited in Foltz, P. W. (1996), Comprehension, coherence and strategies in hypertext and linear text. In: E. Hatch & A. Lazaraton (1991), *The Research Manual: Design and Statistics for Applied Linguistics*. Boston, Massachusetts: Heinle & Heinle.

Johnson, M. & Greatorex, J. (2006). Judging learners' work on screen. How valid and fair are assessment judgements? *Research Matters: A Cambridge Assessment Publication*, **2**, 14–17.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees*. TOEFL Research Report 59. Princeton, NJ: Educational Testing Service.

Kraemer, H. C. & Thiemann, S. (1987). *How Many Subjects: Statistical Power Analysis in Research*. London: SAGE Publications.

Kurniawan, S. H. & Zaphiris, P. (2001). Reading online or on paper: Which is faster? In: *Proceedings of the 9th International Conference on Human Computer Interaction*, 220–222. August 5–10. New Orleans, LA.

Lin, D. (2003). Age differences in the performance of hypertext perusal as a function of text topology. *Behaviour and Information Technology*, **22**, 4, 219–226.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, **3**, 635–694.

Maughan, S. (2001). On-line Teacher Support: A Teachers' Perspective. CIE, UCLES internal report.

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

McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, **39**, 299–312.

McDonald, S. & Stevenson, R. J. (1996). Disorientation in hypertext: the effects of three text structures on navigation performance. *Applied Ergonomics*, **27**, 1, 61–68.

McDonald, S. & Stevenson, R. J. (1998a). Effects of text structure and prior knowledge of the learner on navigation in hypertext. *Human Factors*, **40**, 1, 18–27.

McDonald, S. & Stevenson, R. J. (1998b). Navigation in hyperspace: an evaluation of the effects of navigational tools and subject matter expertise on browsing and information. *Interacting with Computers*, **10**, 2, 129–142.

Murphy, R. (1979). Removing the marks from examination scripts before remarking them: does it make any difference? *British Journal of Educational Psychology*, **49**, 73–8.

Muter, P. & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour and Information Technology*, **10**, 257–266.

Newton, P. E. (1996). The reliability of marking of General Certificate of Secondary Education Scripts: mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

O'Hara, K. (1996). *Towards a typology of reading goals: RXRC affordances of paper project*. Report EPC-1996-107. Available online at: www.lergonome.org/pdf/EPC 1996-107.pdf (accessed 25 August 2005).

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and online documents*. *Proceedings of the Conference on human factors in computing systems (CHI '97)*. 335–342. New York: Association for Computing Machinery.

Price, B. & Petre, M. (1997). *Teaching Programming through Paperless Assignments: an empirical evaluation of instructor feedback*. *Proceedings of ITiCSE '97*. New York: ACM. Available online at: mcs.open.ac.uk/bp5/papers/1997-ITiCSE/ITiCSE%2097-price-petre.pdf (accessed 25 August 2005).

Raikes, N., Greatorex, J. & Shaw, S. (2004). *From paper to screen: some issues on the way*. Paper presented at the International Association of Educational Assessment conference, Philadelphia, June. Available at: www.ucles.org.uk/assessmentdirectorate/articles/confproceedingsetc/IAEA2000NRJGSS (accessed 25 August 2005).

Rayner, K. & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, New Jersey: Lawrence Erlbaum.

Rothkopf, E. Z. (1978). Analyzing eye movements to infer processing styles during learning from text. In: J. W. Senders, D. F. Fisher & R. A. Monty (Eds.), *Eye movements and the higher psychological functions*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Royal-Dawson, L. (2003). Electronic Marking with ETS Software. AQA Research Committee paper RC/219. In: D. Fowles & C. Adams (2005). *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the International Association for Educational Assessment Annual Conference, Abuja, retrieved February 5, 2006, from: www.iaea.info/abstract_files/paper_051218101528.doc

Salmon, G. (2004). *E-moderating. The key to teaching and learning on-line*. London, UK: Routledge Falmer.

Sanderson, P. J. (2001). *Language and differentiation in Examining at A Level*. PhD Thesis, University of Leeds, Leeds.

Segalowitz, G. M., Poulsen, C., & Komoda, M. (1991). Lower level components of reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review*, **8**, 15–30.

Shaw, S. (2005). *On-screen marking: investigating the examiners' experience through verbal protocol analysis*. University of Cambridge ESOL Examinations, Report No 561.

Shaw, S. D., Levey, S. & Fenn, S. (2001). *Electronic Script Management: Report on an exercise held 20, 21, 22 April 2001*. Cambridge: UCLES internal report.

Shaw, S. D & Weir, C. J. (2007). *Examining Writing: Research and practice in assessing second language writing*. Studies in Language Testing No. 26. Cambridge: Cambridge University Press.

Smith, R. N (1992). *Applications of Rasch Measurement*. Chicago: MESA Press.

Smith, B. & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior*, **23**, 3, 1481–1498.

Sturman, L. & Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association for Educational Assessment Conference, 5–10 October 2003, Manchester, UK.

Suto, W.M.I. & Nádas, R. (*in press*). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*.

Twing, J. S., Nichols, P. D., & Harrison, I. (2003). *The comparability of paper-based and image-based marking of a high stakes, large scale writing assessment*. Paper presented at the 29th International Association for Educational Assessment Conference, 7 October 2003, Manchester, United Kingdom.

Wenger, M. J. & Payne, D. G. (1996). Comprehension and retention of nonlinear text: considerations of working memory and material-appropriate processing. *American Journal of Psychology*, **109**, 1, 93–130.

Whetton, C. & Newton, P. (2002). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, 1–6 September 2002, Hong Kong SAR, China.

Williamson, P. (2003). Setting, marking and awarding: the examination process. In: K. Tattershall, J. Day, H. James, D. Gillan & A. Spencer (Eds.), *Setting the Standard*. Manchester: AQA.

Zhang, Y., Powers, D. E., Wright, W. & Morgan, R. (2003). *Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12)*. Princeton, NJ: Educational Testing Service.

Holistic judgement of a borderline vocationally-related portfolio: a study of some influencing factors

Martin Johnson Research Division

Background

Literature suggests a number of background issues that might be pertinent to this area of work. The assessment of a large portfolio of mainly textual evidence demands an assessor to accommodate a great deal of information. It has been suggested that assessors' initial comprehension of a text is an important consideration (Huot, 1990b; Sanderson, 2001). This comprehension process is influenced by the linear nature of the reading process which leads to the gradual construction of a mental representation of the text in the head of the reader (Johnson-Laird, 1983). Another cognitive factor to consider relates to the use of 'generic' phrases in assessment criteria. Oates (2004) argues that these can exact a large cognitive demand on assessors if their use is dispersed across different contexts and/or assessors do not encounter the descriptors very frequently. Finally, it is important to consider the value system within which the reader/assessor is located and which might affect their thinking. Sanderson (2001) suggests that the social context of the assessor is important to consider since it recognises their participation in a community of practice (Wenger, 1998) and constitutes an 'outer frame' for their activity.

It is also important to consider how assessors integrate and combine different aspects of an holistic performance into a final judgement. Most study findings appear to support the suggestion that between-marker reliability is greater for analytic scoring methods, where individual scores are given across multiple dimensions, rather than holistic scoring methods, where a single score is given across multiple dimensions (Breland, 1983; Huot, 1990a; Johnson *et al*., 2001). Laming (2004) argues that this is because linear combinations of individual diagnostic signs have greater accuracy than more strictly holistic judgements because they use an arithmetic basis. Other studies also discuss this problematic issue, suggesting that overall judgement is often based on the cumulative weighting and combination of cues found within a performance and that these weightings might vary (Vaughan, 1991; Einhorn, 2000; Elander and Hardman, 2002).

The recent works of Engeström (2001) and Wenger (1998, 2000) have been very influential in terms of recognising the importance of socio-cultural influences for understanding individual behaviours. This has implications for inter-assessor consistency because it suggests that there is a need to reflect on the role that the social dimension plays in assessment judgements including the potential existence of differing interpretations and standards between assessors.

Investigating socio-cultural influence on assessor consistency has implications for the research method chosen. Whilst socio-cultural theory suggests that human behaviour needs to be understood in the context of the interactions between the characteristics of people and their environments, Rapport *et al*. (2004) characterise 'scientific' knowledge as being independent of time and place with observed variations explained through relevant theory. Popular cognitive research methods, such as Kelly's Repertory Grid (Kelly, 1955) or Verbal Protocol elicitation techniques often conform to this experimental scientific model, focussing on individualised data collection whilst potentially overlooking the influence of the social environment on those elicitation processes. On the other hand, descriptive qualitative methodologies, such as observation and interview techniques, can consider the interaction of both social and individual elements. Bronfenbrenner (1979) argues that understanding might be progressed by uniting the schismatic experimental and descriptive psychological traditions through designing research studies which combine ethnographic and more 'controlled' methods.

This present study attempted to accommodate both of these perspectives by using an integrated approach to data collection. It sought to explore issues of consistent assessor judgement by gathering data about individual assessors' cognitive activity as well as the socio-contextual features in which their practices were undertaken.

Method

This study was set in the context of the OCR *Nationals* in Health and Social Care (Level 2). This qualification was chosen because assessors use an holistic, best fit grading model, organised into a number of Assessment Objectives (AO) to judge portfolios of students' work. Six assessors were involved; four assessors (M1-M4) were Visiting Moderators for the qualification and the others (T5 and T6) were experienced OCR *Nationals* course tutors.

In order to investigate the factors that they attended to during the assessment process the assessors were asked to 'think aloud' whilst they judged a Unit 10 (*preparing to work with people with disabilities*) portfolio which had already been identified as having pass/merit borderline characteristics. This commentary, taken to be a partial record of the features that the assessor attended to during the assessment activity, was transcribed into a verbal protocol and analysed with qualitative text analysis software.

A modified Kelly's Repertory Grid (KRG) interview technique was also used to gather data about different assessors' perceptions of constructs within the same assessment criteria. This activity focussed on the grading criteria for Unit 1 (*preparing to give quality care*). The theory underpinning this method is based on Kelly's model of Personal Construct Psychology (Kelly, 1955), which suggests that individuals possess a constructed version of their world based on their experience. This construction comprises personally held bi-polar mental constructs which can be elicited through KRG techniques. This method asks individuals to verbalise salient differences and similarities between triads of objects or 'elements'. These salient features and patterns anchor ends of bi-polar constructs along which individuals can place other different objects or 'elements'. This method was used to elicit the constructs that assessors

perceived within the grading criteria for each Unit 1 AO. These constructs were then related to their judgements during the portfolio assessment exercise in order to explore whether data about construct elicitation and grading criteria interpretation could shed light on issues of consistent judgement-making.

Qualitative contextual data were collected through observations of three moderation visits to schools and colleges in different parts of England involving three of the assessors in this study. These visits enabled case study evidence to be collected through structured field notes to record details about the different sections of the moderation meetings, the amount and diversity of work covered, and contextual working information. These data also fed into the drafting of questions for the next level of data collection where each assessor was interviewed following the portfolio assessment activity. These semi-structured interviews gathered information about assessors' professional background in order to highlight any potential influences upon their assessment practices.

The final stage of analysis involved the integration of evidence from the different sources of data collection. In the first instance this entailed isolating the salient features identified within the VP and KRG data and cross-referencing them to the features identified in the observation and interview data to identify any linkages and patterns. It needs to be acknowledged that this process contained a subjective quality. It ignored some of the individual micro level linkages that might have been discernible through a more fine grained analysis in order to focus on triangulation at the macro level to identify the larger themes within the data.

Table 1: Frequency of assessor judgements at each grade

		Fail	*Pass/Fail*	*Pass*	*Merit/Pass*	*Merit*
AOs	1			1	1	**4**
	2	1		**3**		2
	3	2		**4**		
	4	1		3		**2**
	5	1	1	**2**	1	
	6	1		**2**	2	

*Bold indicates agreement with original portfolio assessment

Table 2: Mean assessor agreement levels

	M1	*M2*	*M3*	*M4*	*T5*	*T6*
M1	—	0.17	0.8	0.5	0.67	0.25
M2	0.17	—	0.33	0.67	0.67	0
M3	0.8	0.33	—	0.67	0.8	0
M4	0.5	0.67	0.67	—	0.8	0
T5	0.67	0.67	0.8	0.8	—	0.25
T6	0.25	0	0	0	0.25	—

Findings

Although this study was not solely concerned with gathering reliability data, differences between the frequencies between assessors' judgements at different grades during the assessment exercise suggested that there was potential for further investigation of the factors that might have affected their judgements (Table 1).

T5 exhibited the greatest overall degree of agreement with other assessors (Table 2). T6 was the most severe assessor. M3 and M2 had the highest and lowest respective levels of agreement with the most senior assessor (M1).

It is important to acknowledge two potential factors that might have influenced the assessors' judgements: it is possible that the think aloud data collection method might have influenced the assessment process; and two of the assessors (M2 and T5) suggested that they lacked familiarity with the particular unit being assessed since both lacked teaching experience of this particular unit, although they both moderated the unit.

The areas of high shared focus in this study were found around areas of the portfolio that were 'signposted' by textual devices such as clear headings and titles. This search for evidence was itself clearly structured by the grading and KUS (knowledge, understanding and skills) assessment guidance as assessors tended to navigate the portfolio by searching for performance evidence in a similar order. Those assessors who rated Unit 10 AO2 most severely were more likely to attend to features embedded within the text and away from the common areas of attention around the 'signposts', and particularly further on in the portfolio.

There were very clear areas where assessors' comments suggested that they were attending to similar ideas and basing their decisions on similar frameworks. In some Unit 10 AOs it was apparent that fundamental values influenced assessors' practice. In AOs 5 and 6 the dominant influence of 'care values' was evident whilst in AO3 it was 'application'. A 'positive assessment' culture also appeared to pervade the practices of these assessors where they looked to highlight the achievement of the learner. This contrasts with some of the practices identified in other areas of general/academic assessment (Sanderson, 2001; Crisp and Johnson, 2007). These positive assessment practices appear to be underpinned by a strong desire to motivate learners, which was a theme clearly articulated by different assessors during interview. One potential concern that this raises is that assessors might tend to give learners the benefit of any doubt when they are in two minds about the quality of a performance, particularly if schools/colleges fail to prepare their students with appropriate tasks or guidance. KRG analysis also alluded to the presence of shared values through the identification of four 'core' constructs across the different Unit 1 AOs. These constructs were: *application* (4 AOs); *description or account quality* (4 AOs); *sources* (4 AOs); and *example use* (3 AOs). Of these, *application* was notable because assessors consistently weighted it very highly, suggesting it to be a very strong core feature of assessment for these judges.

There was also evidence that assessors' values might have affected their practice in other common ways. Verbal protocol analysis showed that some elements within the grading criteria tended to be attended to more than others, perhaps reflecting the value placed on them by the assessors. Assessors appeared to inherently respect having another competent professional to judge the student's proficiency within a contextualised learning environment. In this study assessors alluded to some of the potential problems that this might lead to, particularly when assessors are not given the right degree of information or where it isn't provided in a useful format. The verbal protocol data also suggested evidence of an assessor using the student performance on the practical

task AO to justify her final judgement for the whole portfolio.

There was also evidence of discrepant practice between assessors. Verbal protocol evidence showed that some assessors adopted a linear strategy to combine several equally weighted factors within AOs, whilst others assigned some performance factors unequal weighting. One example of this was Unit 10 AO4 which contained a third party witness statement suggesting that the student's performance warranted a pass grade. Two assessors appeared to assess this practical task evidence in a linear fashion, balancing it equally alongside other AO evidence, and reaching a 'merit' grade overall. For the other assessors it appears that the witness statement might have been a major influencing factor on their final evaluation which suggested a 'pass' grade overall.

Assessors elicited 131 KRG constructs over the six AOs. The most senior assessor (M1) elicited more constructs on average per AO (7.8) than either the other moderators (4.9) or the tutors (5.0), and t-test analysis showed that this difference was significant ($t = 8.16$, $p < 0.01$). Despite this level of verbalisation the most senior assessor found it difficult to separate these constructs into component aspects across the borderlines, potentially signifying the highly tacit nature of important features of this knowledge.

KRG analysis also identified some potentially problematic issues around lexical interpretation. Some of these clustered around 'construct fusion'. It was possible to find instances where assessors felt that the concepts of 'quality' and 'quantity' had become fused as they progressed through the grade descriptors, such as where descriptors used adjectives relating to the quality of a concept (e.g. *simple* or *basic*) alongside adjectives relating to their quantity or existence (e.g. *some*) (Unit 1: AO1 and AO3). Some assessors also perceived that some qualitative aspects of the descriptors lacked discrimination or appeared to overlap. Assessors sometimes expressed difficulty in separating some of the descriptive qualities within the criteria because the terminology failed to adequately describe differences as they understood them. For example, '*organising information appropriately*' (Unit 1: AO2 pass) might also involve it being '*clear, accurate and detailed*' (Unit 1: AO2 merit), or, assessors might expect a '*basic*' understanding of an issue to be also '*sound*' (Unit 1: AO2 and AO3).This issue also linked to the parallel finding in the interview data where some assessors suggested that they knew where to locate commonly agreed meanings for important words, although the location of this resource varied. This aspect of consistent application, and the potential for misaligned understandings, also resonates with other anecdotal data from the early set up stages of the project which suggested that tutors in schools/colleges sometimes assign their own common 'in house' meanings to descriptor terminology.

The verbal protocol data appear to suggest that assessors might find it difficult to focus on particular performance elements in isolation when reading through work. This highlights a potentially central tension for these vocationally-related assessors who have a strong philosophical attachment to holistic assessment. It is also possible to suggest that holistic assessment might allow assessors to avoid areas of an assessment scheme where there is a lack of clear understanding about the meaning of certain criteria. Although this can lead to better levels of consistency it potentially masks a problem nested within the assessment criteria and which needs to be dealt with.

The observation and interview data identified some key pressures relating to the workloads of moderators. For example, they were under pressure to complete the moderation paperwork during their school/college visit whilst at the same time fostering and maintaining positive links with their hosts to support their ongoing development. These demands are potentially contradictory, with the external validity of the qualification at risk if the balance is not correctly struck.

Another interesting issue found in the interview data was the existence of networks beyond the bounds of this qualification that might have had an effect on assessor practice. Assessors 1 and 3 exhibited the highest levels of inter-assessor agreement in the portfolio assessment exercise and they also shared some common frameworks which did not necessarily overlap with other assessors. These shared frameworks included an understanding that 'evaluation' required 'justification', 'synthesis' acted as a key quality indicator, and the use of a linear rather than a holistic method when accumulating different elements into a final judgement. It might be tentatively suggested that these similarities might have been reinforced by the close connection that these assessors had through their contact through moderation work in another Health and Social Care qualification. Acknowledging the possibility that this external link might overlap into the *Nationals* environment is important because it represents one of the networks (and related tools) that might exist and to which some assessors have restricted access.

Implications

The manner in which the assessors balanced some of the information when reaching a judgement appeared to interact with their underlying values. It could be important for these values, of which 'application' and 'generality and synthesis' appear to be core elements, to be elicited and acknowledged. This might help to undermine the often tacit nature of vocational values and help to promote a common codified framework as a basis on which to discuss interpretations of performance evidence.

There was evidence that some assessors tended to combine performance features in a linear fashion whilst others allowed certain features to dominate their overall judgements. Theory suggests (e.g. Laming, 2004) that the linear method should promote better consistency levels but it is important to explore why some assessors might value particular aspects of performances more than others. Discussion about the appropriate way to balance such features could form an important part of the initial training for assessors new to the qualification and their subsequent moderation visits.

The KRG data suggest that the more experienced assessors (who elicit the greatest number of constructs) might find it most difficult to break down their judgement-making processes. This might represent a challenge to the induction of new assessors.

Concerns about 'construct fusion' require a careful evaluation of the grading criteria to trace the development of constructs through boundaries in order to identify where aspects of concept quality or quantity might overlap. The KRG methodology might be a useful technique for such an activity. A consequence of this process would also be to allow training and moderation visits to draw assessors' attention to this potential problem within the criteria so that they can be aware of it when making judgements. This feature could also factor into any future assessment criteria development programmes.

Consistent lexical interpretation could be further supported by having a clearly referenced resource available for qualification users that defines the meaning of key terminology (e.g. the terms '*range*' or '*simple*' and '*detailed*'). This would reinforce the messages given at training sessions where literal explanations of terminology might be given to new qualification users. This could also be followed up through discussions

around the meanings of key terms during moderation visits.

Assessors sometimes expressed difficulty in separating some of the descriptive qualities within the criteria because, from their perspective, the terminology failed to adequately illustrate differences between the qualities of different performances. This implies that the language used either did not conform to discrete categories or had some overlapping qualities (e.g. '*clear*'/ '*accurate*'/ '*appropriate*'/ '*detailed*' or '*basic*'/ '*sound*'/ '*high*'), that made it difficult for assessors to fit some performance characteristics to the criteria. Although caution needs to be expressed about making assessment criteria more lengthy (Wiliam, 1998; Wolf, 1995), resolving this issue might involve clarifying the values implicit in the descriptor terminology, perhaps through exemplification, and connecting these meanings through effective communication procedures with assessors' expectations about performance quality. This implies a need to engage assessors in discussions about those aspects of language that they feel hinder their ability to discriminate between performances and to use this as an opportunity to arrive at agreed meanings.

References

Breland, H. (1983). *The direct assessment of writing skill: a measurement review*. Technical Report No. 83–6. Princeton, NJ: College Entrance Examination Board.

Bronfenbrenner, U. (1979). *The ecology of human development: experiments by nature and design*. Cambridge, MA: Harvard University Press.

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–962.

Einhorn, H. J. (2000). Expert judgement: some necessary conditions and an example. In: T. Connelly, H. R. Arkes & K. R. Hammond (Eds.), *Judgement and decision making: an interdisciplinary reader*. 2nd edition, 324–335. Cambridge: Cambridge University Press.

Elander, J. & Hardman, D. (2002). An application of judgement analysis to examination marking in psychology. *British Journal of Psychology*, **93**, 303–328.

Engeström, Y. (2001). Expansive learning at work: toward an activity theoretical reconceptualization. *Journal of Education and Work*, **14**, 1, 133–156.

Huot, B. (1990a). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, **60**, 237–263.

Huot, B. (1990b). Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication*, **41**, 210–213.

Johnson, R. L., Penny, J. & Gordon, B. (2001). Score resolution and interrater reliability of holistic scores in rating essays. *Written Communication*, **18**, 2, 229–249.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.

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

Laming, D. (2004). *Human judgment: the eye of the beholder*. London: Thomson Learning.

Oates, T. (2004). The role of outcomes-based national qualifications in the development of an effective vocational education and training system: the case in England and Wales. *Policy Futures in Education*, **2**, 1, 53–71.

Rapport, F., Wainwright, P. & Elwyn, G. (2004). "Of the edgelands": broadening the scope of qualitative methodology. *Journal of Medical Ethics; Medical Humanities*, **31**, 37–42.

Sanderson, P. (2001). *Language and differentiation in Examining at A Level*. PhD thesis, University of Leeds.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, **7**, 2, 225–246.

Wiliam, D. (1998). *Construct-referenced assessment of authentic tasks: alternatives to norms and criteria*. Paper presented at the 24th Annual Conference of the International Association for Educational Assessment, Barbados.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Annotating to comprehend: a marginalised activity?

Martin Johnson Research Division **and Stuart Shaw** CIE Research

Introduction

One of the important premises underlying this article is that the cognitive processes involved in reading can play a significant role in assessment judgements. Although we acknowledge that not all assessments of performance rely on assessors appraising written texts, many tests use written evidence as an indicator of performance. As a result, it is important to consider the role of assessors' comprehension building when reading candidates' textual responses, particularly where candidates are offered a greater freedom in determining the form and scope of their responses.

Crisp and Johnson (2007) note that it is common practice for examiners to annotate scripts when marking. This convention is formalised in the Qualifications and Curriculum Authority (QCA) code of practice (QCA, 2007) which stipulates that a second assessor needs to see any annotations made by a first assessor to gain a full and clear understanding of whether the marking criteria have been applied as intended. Beyond this formalised role, annotation might perform a more general and less formalised function in individual reading comprehension building processes.

Sources (Weiner and Simpson, 2005; Merriam-Webster, 2005) suggest that the definition of the word 'annotation' is to be found in the

15th Century Latin word 'annotare' meaning 'to note or to mark'. The historical importance of the activity is highlighted by Manguel (1997) and Wolfe and Neuwirth (2001) who suggest that it provided a social function, facilitating the sharing of meaning in mediaeval literary cultures. Modern annotation, however, tends to be defined as a discrete activity. Most commonly it is defined as an explanatory note (Weiner and Simpson, 2005), a note added by way of comment or explanation (Merriam-Webster, 2005), a short definition (Nation, 1983), an explanation of word meaning (Pak, 1986), or a critical or explanatory commentary or analysis added to a text (Wiktionary, 2008). Some definitions also allude to the wider impact of annotating on the annotator and any other subsequent reader. Cousins *et al*., (2000) define annotation as a commentary on an object that the annotator intends to be, and the reader interprets to be, separable from the object itself.

This article considers how annotation might influence reader comprehension building at an informal personal level whilst also fulfilling other more formal functions within assessment processes. It goes on to explore how constraining this informal personalised activity might also influence those comprehension building processes. In order to explore how annotation may impact on text comprehension it is first necessary to ascertain what the literature reveals about the various theories and models of reading comprehension.

Models of reading comprehension

Reading is a complex cognitive activity. Attempts to articulate understandings of the reading comprehension process are neither new nor simple (see Huey in Anderson and Pearson, 1988). Prevailing language processing theories offer insights into the mental processes involved in readers' text comprehension when engaging in different types of real life reading. The intricacy of the cognitive processing activities involved in reading are described in varying degrees of detail by Alderson, 2000; Birch, 2007; Cohen and Upton, 2006; Field, 2004; Grabe and Stoller, 2002; Kintsch and van Dijk, 1978; Perfetti, 1999; Rayner and Pollatsek, 1989; and Urquhart and Weir, 1998. Weir and Khalifa (2008) provide a very helpful overview of a range of contributions to the body of theory concerning reading comprehension. Most of the literature cited draws heavily on first language (L1) research and many of the established theories concerning reading comprehension and language processing resonate with current thinking in the fields of cognitive psychology, psycholinguistics, and language assessment.

Developments in reading research over the last century have highlighted significant shifts in the way that the reading process has been perceived: moving from a bottom-up to a more integrated interactive model via a top-down approach.

Bottom-up processing models of reading comprehension require that the reader utilises a range of orthographic, phonological, lexical, syntactic units in order to progress along the scale of linguistic processing. Beginning with recognition of individual letters, followed by words and then sentences, the reader converges on a sense of textual meaning at both a local and global level. Comprehension on a global level relates to propositional understanding (literal interpretation) beyond the level of the text's microstructure and involves the reader's background knowledge along with their ability to identify arguments; recognise central concepts, key details and textual features, such as gist, coherence, cohesion and rhetorical structure. Local comprehension is related to linguistic knowledge (Cohen and Upton, 2006) and takes place on the micro-structural, sentence and clause level. Local comprehension is associated with the understanding of micro-propositions such as word meaning and memory, sentential syntax, and textual details, amongst other things.

In top-down models of processing, comprehension is accomplished through the integration of incoming information with the reader's existing knowledge structures. Propositional meaning, or literal interpretation, is built and developed as readers combine what they encounter in the text with the linguistic, content and cultural knowledge they bring to the text. Thus in the act of reading, readers employ existing schemata to both develop 'meaning representation of the text so far' (Weir and Khalifa, 2008, p.6) and to predict subsequent text.

In the interactive model of reading comprehension, processing takes place in both directions, proceeding simultaneously:

> *Reading involves the simultaneous application of elements such as context and purpose along with knowledge of grammar, content, vocabulary, discourse conventions, graphemic knowledge, and metacognitive awareness in order to develop an appropriate meaning*. (Hudson, 1991, p.83).

Presently, it is widely held that readers construct meaning by processing at different levels concurrently, employing both top-down and bottom-up processing.

Reading comprehension as a metacognitive activity

In their forthcoming volume, *Examining Reading*, Weir and Khalifa present a cognitive processing approach to defining reading comprehension. They identify from the literature within the field of cognitive psychology, certain generic cognitive processes which contribute to the reading process. The cognitive model they use is based on the earlier work of Urquhart and Weir (1998) which expanded Just and Carpenter's (1980, 1987) model and incorporated components from Kintsch and van Dijk (1978, 1983). Central to the model is an understanding of three key constituent features: the *goal setter*; the *processing core*, and the *monitor*. What follows is a very brief description of the role and function of each of these three components. These are considered to be important because annotation may interact with these components and influence key metacognitive functions that facilitate reading comprehension.

The overall goal of reading activity is determined by the *goal setter* which also selects the form of reading which is most likely to realise that goal. Having established a purpose for the reading, the reader is better placed to identify and select the most suitable strategies and determine the type and nature of information which needs to be targeted in the text. Urquhart and Weir (1998) present a helpful matrix in which they identify reading strategies and skills (careful and expeditious) at the local and global levels. 'Strategies' can be thought of as cognisant analytic activities and 'skills' as subliminal, perfunctory abilities (Cohen 1998; Urquhart and Weir, 1998).

The *central processing* element characterises a sequence of reading behaviours. Weir and Khalifa (2008) describe each of these behaviours in detail. Visual recognition, which constitutes the first level of processing, comprises *word recognition* and *lexical decoding*. Word recognition relates to the matching of the form of a word as manifested in written text with a mental representation of the orthographic forms of the language. According to Field (2004), lexical access/decoding, is the 'retrieval of a lexical entry from the lexicon, containing stored

information about a word's form and its meaning'. The next level of processing constitutes an important feature of comprehension. *Syntactic parsing* is concerned with the assembling of words into larger textual units and helps to establish *propositional (core) meaning* at clause and sentence level:

> *Propositional meaning is a literal interpretation of what is on the page. The reader has to add external knowledge to it to turn it into a message that relates to the context in which it occurred.* (Weir and Khalifa, 2008, p.9).

Inferencing, the next higher order level of processing, is a necessary and creative process resulting in the addition of information brought to the text by the reader in an attempt to make the text more meaningful. The reader is now in a position to *build a mental representation* (or model) of the text:

> *... incoming information has to be related to what has gone before, so as to ensure that it contributes to the developing representation of the text in a way that is consistent, meaningful and relevant. This process entails an ability to identify main ideas, to relate them to previous ideas, distinguish between major and minor propositions and to impose a hierarchical structure on the information in the text.* (Field, 2004, p.241).

Creating a text-level structure constitutes the final phase of language processing in which a discourse-level structure is constructed for the entire text.

The *monitor* is a mechanism which provides the reader with feedback regarding the efficacy of the selected reading process. There is a 'symbiotic' relationship between the monitor and goalsetter in that the monitoring process is reliant upon the decisions taken with regard to the type of reading and, therefore, the monitor is triggered in accordance with the goalsetter. Thus each component acts as a metacognitive device that mediates among the reader's range of processing skills and knowledge sources.

Thus the reading process can be thought of as an interaction of the reader's conceptual abilities and process strategies, and their language knowledge and content knowledge.

Annotation as a support for reading comprehension

Anderson and Armbruster (1982) and O'Hara (1996) identify a number of written support activities that are commonly associated with reading. This evidence has led some (O'Hara and Sellen, 1997; Marshall, 2001) to observe that such activities can often operate concurrently with reading activity, frequently being seamlessly integrated with reading activity, and habitually being unselfconsciously generated by the annotator. Wolfe and Neuwirth (2001) also cite a study by Adler *et al*. (1998) which found reader annotation activity occurring in conjunction with reading activity more than 25% of the time, with an additional 22% of annotations being made on separate documents from the reading source document. It appears that the reason for the existence of such practices could relate heavily to the cognitive processes involved with reading comprehension. This observation is supported by research evidence which has found that the complexity of a reading task influences reading performance (Mayes *et al*., 2001; Just and Carpenter, 1987). Weir (2005) theorises the cognitive complexity of such processes. One particular area of the central processing core appears to be of specific interest when discussing annotation practices. The process of building a mental model of a text involves reader self-monitoring, which in turn involves the use of working memory. It appears that annotation might perform an important function in mediating reader workload and enhancing comprehension.

There is a body of research which explores how annotating might actually support comprehension building processes. Hsieh *et al*. (2006) highlight evidence from Hartley and Davies (1978) that annotating facilitates textual encoding during the reading process. Textual encoding involves the basic perceptual process of converting a sensory input into subjectively meaningful experience. This encoding process plays a central role in reading comprehension. Weir and Khalifa (2008) outline how the central processing core involves a reader building a mental model of a text through integrating visual textual information with their world knowledge. Annotating might play an important role in this integration. The reason for this might be explained by the way that annotating involves the active integration of a reader's present understanding with new information encountered within the text. Sometimes this might involve the reader paraphrasing or elaborating on textual information in the form of an annotation.

Another important aspect of encoding also involves spatial encoding. Piolat *et al*. (1997) argue that a number of research findings are consistent with the idea that spatial encoding occurs during reading activity and that this is an integral part of a reader building a material representation of the location of textual information. In other words, reading is a spatial activity with the reader's eyes moving from one fixation location to the next to pick up spatially distributed visual information and processing positional information. This interpretation is corroborated by Fischer (1999) who argues that there is both direct and indirect evidence to suggest that memory is used to process information about spatial attributes of texts during reading. This work implies that the act of reading involves the mental spatial tagging of ideas and concepts in a text rather than the tagging of the location of words alone. Such research evidence also reinforces the postulation by Kennedy (1992) of a 'spatial coding hypothesis'. This hypothesis intimates that readers consider texts to behave as physical objects which provide the reader with spatial code in addition to lexical information. A tangible outcome of this hypothesis is demonstrated in studies that highlight how reader information recall correlates positively with increased reader annotation (Hartley and Davies, 1978; Hartley, 1983; Khan, 1994).

Annotating might also perform an important metacognitive function during reading. According to Weir and Khalifa (2008) self-monitoring is a complex metacognitive operation that provides the reader with feedback about the success of their reading processes. McMahon and Dunbar (2003) investigate tools that might support comprehension monitoring and suggest that annotation might support this function. This phenomenon was also observed in a study by Crisp and Johnson (2007). Examiners involved with assessing longer textual answers were observed making annotations whilst reading and they suggested that these annotations provided them with an individual checking function or a means to communicate with themselves about the text being read. This also appears to link to research observations which suggest that annotations might support such a metacognitive function by aiding working memory in a retrospective manner. Marshall (1997) reported that readers' annotations were being used as a visible trace of the reader's attention, especially when the text was in a protracted narrative form. Marshall suggested that these annotations could act as place

markings that subsequently aided the annotator's memory. This suggests that annotation can function as a storage bank of information external to individual working memory.

It might be important to reflect on the idea that annotating practices are potentially highly individualistic in character. Crisp and Johnson (2007) and Shaw (2008) found that some examiners were prone to using annotations idiosyncratically despite the clearly defined expectations of the mark schemes to which they were working. This relates to the view that annotations might be seen to represent the point of convergence between a reader's current knowledge and the propositional aspects of a text that they are encountering at a given time. It is reasonable to assume that the tangible outcome of that encounter would be particular to that situation. This reflexive quality might be important given research which suggests that highly individualistic note-taking can facilitate better information encoding and storage external from working memory (Hartley and Davies, 1978; Hartley, 2002).

Annotation and assessment

This article has highlighted the potential role of annotating on reading comprehension processes. Moreover, this activity is essentially an informal and potentially highly individualistic activity, influenced by the interaction of a variety of particular factors at a given time.

There have been relatively few studies that have looked at annotating activities in the context of educational assessment, but the limited literature suggests that annotations perform additional functions which are specifically linked to the context of large scale examination processes. Two recent studies at Cambridge Assessment suggest that assessor annotation performs a number of functions beyond supporting reader comprehension.

Crisp and Johnson (2007) report evidence of examiner annotations serving two distinct functions. The first function was to facilitate examiner judgements. The study found that examiners found annotating to be particularly useful for reinforcing their comprehension of protracted texts. The second function was a justificatory one, where annotations communicated the reason for a judgement to other assessors within the system. In this sense annotating supported the confidence of examiners to complete their marking in the knowledge that others would be aware of the reasoning behind their assessment judgements. This confidence factor also parallels the findings of Shaw (2005) who found that examiners used annotations to investigate their own marking consistency. Shaw observed that annotations were used by examiners to provide an efficient means to confirm, deny or reconsider their marking standards both within and across candidates, thereby reassuring the examiners throughout the marking event.

The findings of these studies suggest that annotating activities in large scale assessment systems might be influenced by competing demands beyond the basic requirement to support individual examiner comprehension building. The reason for this might be explained by the accountability concerns attached to large scale assessment, and the related objective of maximising transparent communication within the assessment system.

The case of large scale examinations in the UK provides a useful context within which to discuss this issue. The accountability agenda that pervades education has led to public examinations being the most widely used performance indicators for educational success. The scale of this exercise requires measures to ensure that the examination system functions in a fair and robust manner. The Code of Practice (QCA, 2007), produced by the examination regulatory authorities in England, Wales and Northern Ireland, outlines procedures that awarding bodies should follow to ensure that examinations are developed and administered within transparent and accountable structures. There is an unambiguous emphasis on clear communication channels between examiners of different seniority to facilitate effective monitoring. Williamson (2003) comments that this function is all the more important in an expanding examinations system, such as that of the UK. Annotations have an important communicative role in this quality control process.

The importance of justificatory annotating undoubtedly influences examiners' practices. This is visibly demonstrated by the extent to which annotating activity is documented in assessment guidelines. Besides the guidance given in mark schemes, at marker standardisation and at coordination meetings, expectations are set out by the QCA. These formalised arrangements state that internal coursework assessments and associated assessment criteria must indicate how credit has been assigned, and that therefore 'Internal assessors are required to annotate the coursework, clearly showing how the marking criteria have been applied' (QCA, 2005, p.19). The most recent QCA code of practice (QCA, 2007) also requires that principle moderators must 'compile exemplar work, annotated to show how the assessment criteria are to be applied' (p.9), in order to ensure that the standards of the unit or component are maintained and consistent with the unit specification and assessment criteria. The code of practice also requires awarding bodies in the UK to 'continue to mark and annotate all scripts in accordance with good practice recognised by the regulatory authorities' (QCA, 2007, p.49).

Wolfe and Neuwirth (2001) suggest that annotations can fulfil a variety of functions, although it appears that in different contexts some functions may dominate others. Wolfe and Neuwirth suggest that annotations can facilitate reading and later writing tasks; eavesdrop on the insights of other readers; provide feedback to writers or promote communication with collaborators; and call attention to topics in important passages. The emphasis in formalised assessment discourse about annotation practices appears to accentuate the functions of eavesdropping on the insights of readers and promoting collaboration with others rather than reading facilitation. The consequence of this in assessment practice is that certain annotation conventions come to be considered acceptable and become 'expected' practice in order to promote transparency and consistency amongst examiners. As a result, it might be argued that the prioritisation of the accountability function could lead to the demotion of the comprehension building function which might rely more on flexible and individualistic annotation practices.

Literature suggests that the mode in which a text is presented, either on paper or on screen, also represents another area where systematic pressures come to bear on annotating practices. O'Hara and Sellen (1997) argue that mode can affect reader annotation in a number of ways. One major concern is the degree of physical effort required to annotate in one mode compared with another. They suggest that making paper-based annotations is a relatively effortless procedure and, as a consequence, it factors automatically into the meaning construction process during reading. In contrast, computer-based annotation practices can be impeded by the availability of authentic annotation tools. Keyboards might influence annotating behaviour because they do not accommodate many of the types of mark that readers choose to use when working on paper, therefore making the process less genuine and positively affecting the cognitive demand on the reader.

Again, there is limited empirical literature on modal influence on assessor annotation practices. In one study involving higher education instructors, Price and Petre (1997) present a mixed picture of modal influence. They found that all the instructors in their study had different marking styles on paper and on screen, but that the extent of this modal influence varied between markers. They also found that the annotations used in paper and electronic marking were different, with underlining, circling and highlighting being used less in electronic marking than on paper, despite their availability. Despite these differences, Price and Petre conclude that technology did not impair the assessment practices, and more specifically the annotating processes, for all markers. Other studies suggest emotive and physical dimensions in relation to computer annotating. Greatorex (2004) reports teacher frustration when moderating electronic portfolios. Her study highlights the difficulties that teachers experienced when annotating candidates' work directly, with teachers reporting that there would have been more annotation if portfolios had been presented on paper. Shaw (2008) observes that marker concentration might be adversely affected when assessing on-screen. Not being able to replicate paper and pen practice when applying annotations was a predominant concern amongst trial participants in his study. Participants generally perceived on-screen marking to be physically more demanding than paper marking. Moreover, marking over prolonged periods engendered mental and physical fatigue with the physical process of selecting and applying annotative tools on-line being demanding.

Conclusion

Crisp and Johnson (2007) have suggested that one of the two functions annotations serve is justificatory and that annotating might have an important communicative role in the quality control process in terms of accountability. Annotating has a particular role in assisting with transparent communication between different markers (Williamson, 2003). Accountability is widely recognised to be a multifaceted and complicated concept (Day and Klein, 1987) and 'assumes the requirement to answer to the broader social community' (Kogan, 1986). In an educational context, examination boards offering high-stakes assessments are required to account for or justify certain assessment actions and behaviour for a range of potential community stakeholders. Thus, the notion of accountability is closely related to responsibility, as those who have been given responsibilities – the assessment practitioners – are asked to account for their assessments. If the conventional accountability processes are influenced by the introduction of a new, computer-assisted assessment medium then both the reliability of test scores and the validity of the assessments are potentially compromised.

By bringing together literature about linguistics and annotation practices, both empirical and theoretical, this article suggests that a critical link exists between annotating and reading activities. Moreover, an important aspect of this relationship is associated with reader comprehension building. It is perhaps significant that empirical study into annotating in assessment contexts is very limited and this helps to explain why the extent to which annotating candidate responses influence or affect assessor comprehension is neither known nor fully understood. This is an important observation since the arguments advanced in this article suggest that such an influence is tangible.

Through making the different functions of annotation explicit the intention of this article is to primarily amplify the importance of the impact of annotating on assessor comprehension. It is also intended that this function be clearly understood in relation to the other accountability and transparency purposes that currently influence how annotation is used in large scale assessment systems. It is also worth noting that this issue has potentially important consequences for ongoing debates about on-line assessment both within Cambridge Assessment, through the current series of on-screen marking trials, and beyond. It is hoped that this article can make a positive contribution to this area of discussion.

References

Adler, A., Gujar, A., Harrison, B. L., O'Hara, K., & Sellen, A. (1998). *A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices*. Proceedings of CHI '98, 241–248.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Anderson, R. C. & Pearson, P. D. (1988). A Schema-theoretic View of Basic Processes in Reading Comprehension. In: P. L. Carrell, J. Devine & D. E. Eskey (Eds.), *Interactive Approaches to Second Language Reading*. Cambridge: Cambridge University Press.

Anderson, T. H. & Armbruster, B. B. (1982). Reader and text-studying strategies. In: W. Otto, & S. White (Eds.), *Reading Expository Material*. London: Academic Press.

Birch, M. (2007). *English L2 Reading: Getting to the Bottom*. Mahwah: Lawrence Erlbaum Associates.

Cohen, A. (1998). *Strategies in Learning and Using a Second Language*. London: Longman.

Cohen, A. D. & Upton, T. A. (2006). *Strategies in Responding to the New TOEFL Reading Tasks* (Monograph No. 33). Princeton, NJ: Educational Testing Services.

Cousins,S. B., Baldonado, M. & Paepcke, A. (2000). *A Systems View of Annotations*. Technical Report P9910022. Palo Alto, CA: Xerox/Palo Alto Research Center.

Crisp, V. & Johnson, M. (2007). The Use of Annotations in Examination Marking: Opening a Window into Markers' Minds. *British Educational Research Journal*, **33**, 6, 943–961.

Day, P. & Klein. R, (1987). *Accountabilities: Five Public Services*. London: Tavistock.

Field, J. (2004). *Psycholinguistics: the Key Concepts*. London: Routledge.

Fischer, M. H. (1999). Memory for Word Locations in Reading. *Memory*, **7**, 1, 79–118.

Grabe, W. & F.L. Stoller (2002). *Teaching and Researching Reading*. London: Longman.

Greatorex, J. (2004). *Moderated E-portfolio Project Evaluation*. Cambridge: University of Cambridge Local Examinations Syndicate.

Hartley, J. (1983). Note-taking Research: Resetting the Scoreboard. *Bulletin of the British Psychological Society*, **36**, 13–14.

Hartley, J. (2002). Notetaking in Non-academic Settings. *Applied Cognitive Psychology*, **16**, 559–574.

Hartley, J. & Davies, I. K. (1978). Note-taking: A Critical Review. *Innovation in Education and Teaching International*, **15**, 3, 207–224.

Hsieh, G., Wood, K. R. & Sellen, A. (2006). *Peripheral Display of Digital Handwritten Notes*. Proceedings of the Conference on Human Factors in Computing Systems, Montreal, Quebec, 2006.

Hudson, T. (1991). A Content Comprehension Approach to Reading English for Science and Technology. *TESOL Quarterly*, **25**, 1, 77–104.

Just, M. A. & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, **87**, 4, 329–354.

Just, M. A. & Carpenter, P. A. (1987). *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon.

Kennedy, A. (1992). The Spatial Coding Hypothesis. In: K. Rayner (Ed.), *Eye Movements and Visual Cognition*. New York: Springer-Verlag.

Khan, F. (1994). *A Survey of Note-taking Practices*. HP Labs Technical Reports HPL-93–107.

Kintsch, W. & van Dijk, T. A. (1978). Toward a Model of Text Comprehension and Production. *Psychological Review*, **85**, 363–394.

Kogan, M. (1986). *Educational Accountability: An Analytic Overview*. London: Hutchinson.

Manguel, A. (1997). *A History of Reading*. London: Flamingo.

Marshall, C.C. (1997). *Annotation: From Paper Books to the Digital Library*. Proceedings of the Second ACM International Conference on Digital Libraries; Philadelphia, Pennsylvania.

Marshall, C.C. (2001). *The Haunting Question of Intelligibility*. Paper presented at the ACM Conference on Hypertext and Hypermedia, Aarhus, August 14–18, 2001.

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.

McMahon, M. & Dunbar, A. (2003). Mark-UP: Facilitating Reading Comprehension through On-Line Collaborative Annotation. In: N. Smythe (Ed.), *Digital Voyages*. Proceedings of the Apple University Consortium Conference, Adelaide, South Australia, September 28 – October 1, 2003.

Merriam-Webster (2005). *Collegiate® Dictionary, Eleventh Edition*. Springfield, MA: Merriam-Webster.

Nation, I. S. P. (1983). *Teaching and Learning Vocabulary*. Wellington: English Language Institute, Victoria University.

O'Hara, K. (1996). *Towards a Typology of Reading Goals*. Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996–107. Cambridge: Rank Xerox Research Centre.

O'Hara, K. & Sellen, A. (1997). A Comparison of Reading Paper and Online Documents. In: S. Pemberton (Ed.), *Proceedings of the Conference on Human Factors in Computing Systems*. New York: Association for Computing Machinery.

Pak, J. (1986). *The Effect of Vocabulary Glossing on ESL Reading Comprehension*. Unpublished manuscript. University of Hawaii at Manoa.

Perfetti, C. A. (1999). Comprehending Written Language: A Blueprint for the Reader. In: C.M. Brown & P. Hagoort (Eds), *The Neurocognition of Language*. Oxford: Oxford University Press.

Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of Screen Presentation on Text Reading and Revising. *International Journal of Human-Computer Studies*, **47**, 565–89.

Price, B. and Petre, M. (1997). *Teaching Programming through Paperless Assignments: An Empirical Evaluation of Instructor Feedback*. Milton Keynes: Centre for Informatics Education Research, Open University.

QCA (2007). *GCSE, GCE, GNVQ and AEA Code of Practice*. London: QCA.

QCA (2005). *A Review of GCE and CSE Coursework Arrangements*. London: QCA.

Rayner, K. & Pollatsek, A. (1989). *The Psychology of Reading*. Englewood Cliffs, NJ: Prentice Hall.

Shaw, S. (2005). *On-screen Marking: Investigating the Examiners' Experience through Verbal Protocol Analysis*. Internal ESOL Validation and Research Report.

Shaw, S. (2008). *On-screen Marking of Extended Writing: Towards an Understanding of Examiner On-line Assessment Behaviour*. Internal CIE Research Report.

Urquhart, A. H. & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice*. London: Longman.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Weiner, E. S. C. & Simpson, J. A. (Eds.) (2005). *Compact Oxford English Dictionary*. Oxford: Oxford University Press.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. & Khalifa, H. (2008). A Cognition Processing Approach Towards Defining Reading Comprehension. *Research Notes 31*, February 2008, 2–10.

Wiktionary (2008). http://en.wiktionary.org/wiki/annotation

Williamson, P. (2003). Setting, Marking and Awarding: The Examination Process. In: K. Tattershall, J. Day, H. James, D. Gillan & A. Spencer (Eds.), *Setting the Standard*. Manchester: AQA.

Wolfe, J. L. & Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *Journal of Business and Technical Communication*, **15**, 3, 333–371.

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

Gill Elliott Research Division

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 – 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:

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.

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.

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

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.

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 definition and taxonomy for Cambridge Assessment

Beth Black Research Division, **Joe Chislett, Anne Thomson**, **Geoff Thwaites and Jacquie Thwaites**

Introduction

The main aim of this research activity was to create a Cambridge Assessment definition[1] and taxonomy[2] for Critical Thinking.

There are a vast number of Critical Thinking definitions in the literature (e.g. Ennis, 1996; Fisher and Scriven, 1997; Paul, 1992), which are highly varied and often multi-faceted. The construct of Critical Thinking is hotly debated, with a number of key battlegrounds. The implications of such differing conceptions reach out beyond academic journals. They impact upon educationalists in a number of practical ways, such as devising the best training or delivery model for Critical Thinking; designing and delivering valid assessments which are authentic and which nurture good Critical Thinking skills in students.

For these reasons, and others listed below, Cambridge Assessment aspired to have a definition of its own:

Cambridge Assessment as the expert

Cambridge Assessment has 20 years of experience in testing Critical Thinking, unrivalled by any other body within the UK. In order to capitalise upon this experience, it seems sensible to have a definition, or clear sense of the construct that we say we are measuring, so we can be sure that our measures are valid and that we are making valid inferences from these assessments.

Coherence

It is important that, across Cambridge Assessment's existent Critical Thinking offerings, there is a coherent understanding of the usage of the term and the construct being measured. This should also be true of any assessments or qualifications developed in the future.

Currently, Cambridge Assessment has five, long term, extant products (see Figure 1): BMAT, TSA, CIE Thinking Skills AS/A level, OCR AS/A Level Critical Thinking and OCR AEA Critical Thinking, all of which share a common ancestor, namely MENO. However, each of them has a slightly different evolutionary history, tests differing aspects and subsets of Critical Thinking, and is used for different purposes and candidate types.

Additionally, there is a newer qualification, namely CIE's H2 Knowledge and Inquiry, which includes a Critical Thinking paper. This is less obviously a descendent of MENO, though it does necessarily involve analysis and evaluation of arguments. Equivalent to A-level, it was developed specifically for Singapore's stronger candidates in order to enhance skills needed for university.

1 Definition: 'stating the precise nature of a thing'

2 Taxonomy: a term, now commonly borrowed from the biological sciences meaning 'dealing with the description, identification, naming, and classification of organisms'

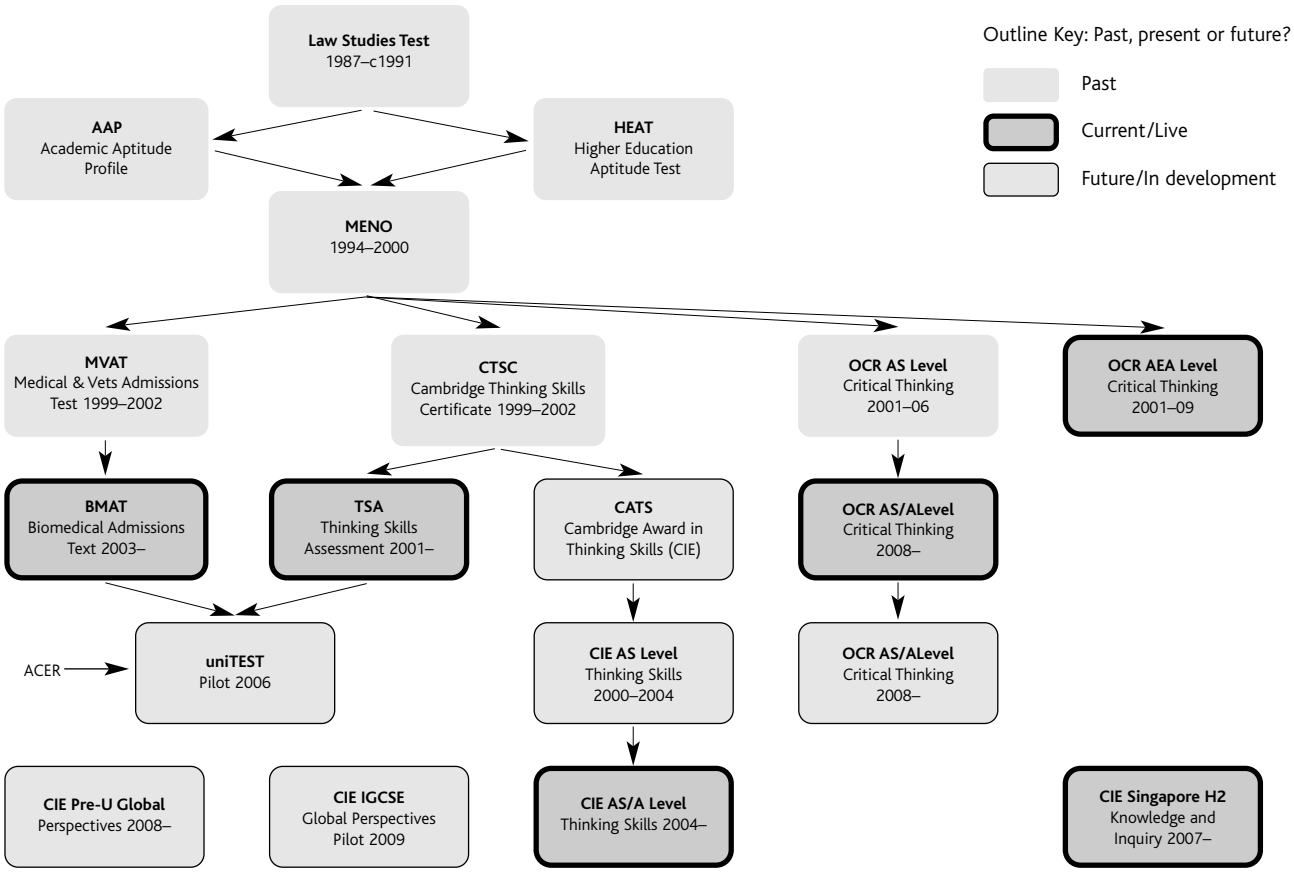


Figure 1: Family Tree of Cambridge Assessment Critical Thinking products

Future Critical Thinking assessments

Another pressing need for a definition relates to the fact that nothing stands still in the world of assessment. A number of new Critical Thinking products are in development. The CIE Thinking Skills specification is altering its scheme of assessment from summer 2008 and OCR has had a new specification accredited (H052/H452) for teaching from September 2008. But more dramatically, a new generation of tests and qualifications is in development. The CIE Pre-U Global Perspectives qualification is one high-profile example. This will be an obligatory unit for those wishing to gain the Pre-U diploma, and contains Critical Thinking elements. Whilst possibly a more applied context than other Cambridge Assessment offerings, this will bring a particularly international dimension to Critical Thinking. CIE is also developing an IGCSE in Global Perspectives, and whilst nothing in the specification is actually called Critical Thinking, there are resonances of Critical Thinking in the pilot specification (e.g. in terms of 'reasoned responses' or 'engaging in enquiry').

Another example is uniTEST, a selection test under development, which is designed to be a general university admissions test with a widening participation agenda. Its Critical Thinking (or Critical Reasoning) items are presented as a middle ground between abilities used in arts/humanities and those used in maths/science.

It is less obvious exactly how these newer products fit into the family tree, and which products are their immediate predecessors. Nonetheless, the developers of many of these new qualifications have also been involved in existent qualifications and so some sort of common understanding of the nature of Critical Thinking is transmitted implicitly.

Looking further ahead, it is likely that the number and nature of Cambridge Assessment tests and qualifications will continue to change and evolve and therefore, for the purpose of coherence of new and future products, it is vital that there is a Cambridge Assessment definition of Critical Thinking. Quite possibly, in years to come, any definition may need to be reviewed in the light of the natural evolution and development of the discipline. Nonetheless, a definition would still have a lifespan useful for the guidance for any development work.

Perceptions of Critical Thinking

Perceptions of Critical Thinking are highly varied and not always based on an informed understanding of the identity and nature of Critical Thinking. This is hardly surprising when academic perceptions and definitions are so multitudinous (for a short summary, see Black, 2007), with philosophical definitions at odds with psychological ones, some focussing more upon skills whilst others emphasise dispositions, and so on. There is also much discussion about what *is* versus what *isn't* Critical Thinking. The outer edges or the fringes of the discipline are not always clear, with much variety in terms of exclusivity or inclusivity of definitions.

Certainly, in terms of size of candidate entry, Critical Thinking AS/A level could be said to be 'popular' in schools: it became OCR's biggest A-level in 2005–6[3], and the fastest growing A-level in the UK in 2007. Within schools, however, teachers hold mixed perceptions of the value of Critical Thinking. At one end of the spectrum some teachers perceive

3 Cambridge Assessment Group Annual report 2005–6 accessed at http://www.cambridgeassessment.org.uk/ca/digitalAssets/110764_Cambridge_Assessment_Group_Annual_Report_2005–2006.pdf on September 12th 2007

Critical Thinking as the 'holy grail' of education, as vital in developing rational argument and reasoned thinking, whilst at the other end teachers (erroneously) see it as something more akin to General Studies. Undoubtedly, there are also a number of teachers who have only superficial acquaintance with the discipline and thus have only a limited idea of what it entails. It is not surprising, therefore, that universities have different policies on the value of Critical Thinking for admissions. For example, some universities do accept Critical Thinking AS/A level as part of their main offer, whereas others look upon it favourably as an additional extra, but will not accept it as part of its main offer.

Still, whatever and however people perceive Critical Thinking, there is evidence that students who take Critical Thinking AS level do better in their other A-levels than those who do not take Critical Thinking (Gill and Black, *in prep*).

Cambridge Assessment, with all of its collective expertise, is in a unique position to respond to the issues identified above and therefore contribute to the long-term integrity and success of its Critical Thinking products.

Method

In the first instance, in December 2006, a large one-day meeting was convened, comprising Cambridge Assessment personnel with responsibility for the various Critical Thinking tests and qualifications, as well as a number of Critical Thinking experts who have had involvement with Cambridge Assessment as item writers and/or senior examiners. At this meeting, the topics for a semi-structured discussion included whether a Cambridge Assessment definition and taxonomy for Critical Thinking were desirable and possible. The participants were unanimous in wanting a definition, and broadly consensual regarding the need for a taxonomy. Various existent definitions of Critical Thinking were considered during this meeting.

Overall, the recommendation from the meeting was that a smaller group of three or four experts should be charged with the task of developing both a definition and taxonomy. It is this activity, which took place over four days in October 2007, which forms the basis for this article.

The experts

The expert panel comprised four Critical Thinking experts, all of whom have worked for Cambridge Assessment in examining and/or item writing and/or specification development in this area. They were chosen in consultation with representatives of Business Streams. The guiding principle in selecting these experts was to have good coverage across existent qualifications and tests (see Table 1 below), as well as to have a range of experience of Critical Thinking (academic, school teaching etc).

These individuals were chosen also for some specific qualities or experience. For example, one of the panel members is commonly regarded as one of the leading UK Critical Thinking experts. Another expert was chosen not only for Critical Thinking knowledge, but also expertise in Problem Solving, and to aid the panel in its consideration of the 'outer edges' of Critical Thinking, that is, those 'higher-order thinking skills' which are not Critical Thinking. Another panel member has been involved with Critical Thinking AS since its beginning, was a member of QCA's Critical Thinking Advisory Group (which, amongst other things, was responsible for QCA's definition), and has experience of teaching a variety of candidate types (from under-achieving to gifted and talented). The fourth has a background in Philosophy and has established his expertise in Critical Thinking in teaching, item writing and being a senior examiner. Between the four experts chosen, there was an aggregate of 57 years of experience in Critical Thinking and six published books.

The definition and taxonomy development took place under the guidance of a member of the ARD evaluation team, who has particular research interests in Critical Thinking.

Tasks for the four-day meeting and organisation of time

The experts were asked to:

- derive a Critical Thinking definition
- derive a Critical Thinking taxonomy
- as far as possible, map Cambridge Assessment qualifications against the taxonomy
- identify skills closely related to Critical Thinking but which are not considered to be Critical Thinking.

The meeting took place over four consecutive days – October 3rd to October 6th 2007. The beginning of the four days was marked by a one-hour plenary session with the relevant CIE, OCR and Cambridge Assessment representatives in order for them to raise construct and definitional issues pertinent to their particular products.

For the main part of the four days, it was deemed to be more productive to allow the experts to decide how to proceed, but offering them three alternative approaches.

The top down approach, working sequentially to derive first a definition as a group, then a taxonomy, followed by the mapping exercise, might be considered the purist's approach, in that the definition is derived before and independent from a consideration of the products. However, an entirely pure approach in this respect may not be achievable: naturally, for the experts, their working knowledge of their products (see Table 1 above) is implicit and bound to inform any work on the definition.

The bottom-up approach involves considering the Cambridge Assessment products in some detail before deriving a definition. In one

Table 1: The four experts and coverage of Cambridge Assessment products

	CIE Thinking Skills	*OCR AS/A Critical Thinking*	*OCR AEA Critical Thinking*	*BMAT*	*TSA*	*uniTEST*	*CIE Singapore H2 Knowledge and Inquiry*
Expert A		✔	✔				
Expert B		✔		✔	✔		✔
Expert C	✔		✔	✔	✔	✔	
Expert D	✔			✔	✔	✔	

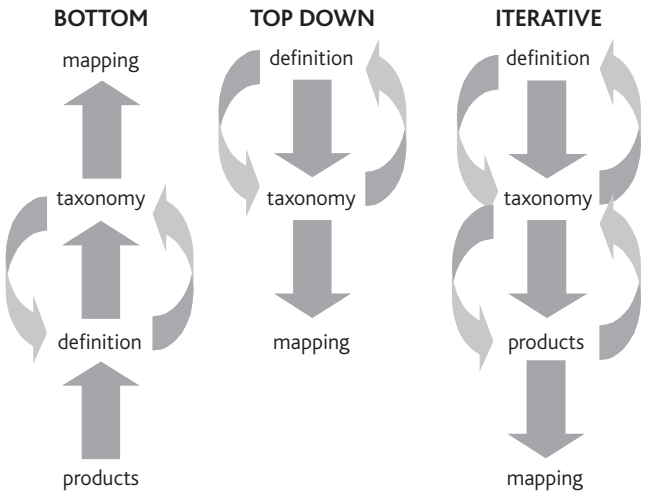


Figure 2: Alternative approaches offered to the panel for the process of determining the definition and taxonomy.

sense, this would be putting a framework around what we have already got, the products themselves providing the driving force for the activity. In other words, the bottom-up process might result in an overly self-confirmatory definition and taxonomy. However, this approach would have an advantage of 'reminding' the panel of (valid) aspects of Critical Thinking.

The iterative approach suggested is based upon the top-down model, where activities logically proceed from the definition. However, this model builds in a capacity to revisit and ultimately refine one step in the light of decisions about another step (as in Figure 2).

Unanimously, the experts chose to adopt the iterative approach. This proved a fruitful approach as, on occasion, the mapping exercise challenged the current version of the taxonomy: for example, the panel questioned whether one sub-skill should be presented as two separate sub-skills, or, conversely, whether two sub-skills were, in reality, inseparable and should be conflated.

There was a range of supporting materials and reference points to draw upon, including many existent Critical Thinking definitions. In particular, experts were guided towards the QCA definition of Critical Thinking (because it was derived in the UK and favoured by the one-day December meeting) and the Facione taxonomy[4] (1990).

During the course of the meeting, it was also decided that the Cambridge Assessment definition should be accompanied by an explication or rationale. The purpose of this is to explain or clarify the intended meaning or choice of words or emphasis contained within the definition. It captures some of the lengthy consideration around the table during the four days and is really intended as a guide for users of the definition. Similarly, the expansion of the taxonomy is again to provide guidance and clarification.

The panel also mapped all Cambridge Assessment products against the taxonomy. All the assessments were mapped by all four panellists. Consensus was achieved through discussion. For this part of the activity, participants had specifications, example exams or tests (usually, the most recent), and where possible, actual examples of student work.

4 Facione's taxonomy was derived by the Delphi method, along with a definition, using a panel of 46 experts. Undoubtedly, Peter Facione's ambitious project to arrive at a definition through expert consensus (Facione 1990) was an attempt to achieve greater harmony amongst Critical Thinkers (in North America at least). However, perhaps the main drawbacks of the Facione definition are its length and its over-inclusivity.

Finally, the definition, taxonomy, rationale and mapping documents were distributed to the relevant subject officers/product managers etc. Some small changes were made (though none to the definition) and the work was very positively received.

Outcomes

The Cambridge Assessment definition of Critical Thinking

> Critical Thinking is the analytical thinking which underlies all rational discourse and enquiry. It is characterised by a meticulous and rigorous approach.
>
> As an academic discipline, it is unique in that it explicitly focuses on the processes involved in being rational.
>
> These processes include:
> - analysing arguments
> - judging the relevance and significance of information
> - evaluating claims, inferences, arguments and explanations
> - constructing clear and coherent arguments
> - forming well-reasoned judgements and decisions.
>
> Being rational also requires an open-minded yet critical approach to one's own thinking as well as that of others.

Rationale/explication of the Cambridge Assessment definition of Critical Thinking

The definition strongly equates Critical Thinking with rationality. Thus, in one sense, Critical Thinking (CT), as an *activity*, is ubiquitous: all rational discourse and enquiry involves the activity and application of CT. Both formal (subject domains across the science-humanities divide) and informal (every day) rational discourse and enquiry rely upon analytical and reasoned thought.

The definition highlights that one of the main features of CT is that it is analytical. Many of the processes of CT rest upon the ability to be analytical; to be able to dissect arguments and information.

Good Critical Thinking is exemplified when the thinking is rigorous and meticulous. That is to say that CT is not passive, automatic, spontaneous or reactive in manner, but is active, careful and thorough.

Whilst CT, as a form of thinking, can be acquired and exercised through incidental exposure in one's general educational experience, the reference to CT as an academic discipline acknowledges that this is a skill which can be explicitly and purposefully learnt and taught. CT comprises a number of processes involved in being rational. These processes are often implicit, hidden or tacit. Studying CT makes these processes unconcealed and explicit. Therefore, whilst a person who has had an absence of any overt CT teaching might still be equipped with a range of CT skills, explicit teaching of CT can introduce awareness or increase proficiency in the processes involved in being rational. The value of the discipline is that it can be applied in all contexts in which reasoning occurs or should occur.

CT emphasises processes – hence the inclusion in the definition of five of the most significant of the many processes of rationality – which encompass the skills and sub-skills outlined in the taxonomy.

Taxonomy with expansion

Skill/process	*Sub-skills/processes*	*Expansion*
1 Analysis	A Recognising and using the basic terminology of reasoning	E.g. argument, reasons, conclusions, analogy, inference, assumptions, flaws. This skill underpins most critical thinking skills.
	B Recognising arguments and explanations	Recognising argument is a fundamental sub skill in Critical Thinking. (An argument is defined as one or more reasons offered in support of a conclusion). Being able to distinguish between argument and non-argument as well as between argument and explanation.
	C Recognising different types of reasoning	Recognising that arguments use different types of reasons, e.g. common knowledge, statistics, conditional statements, scientific data, ethical principles etc. More advanced recognition will include recognising different forms of argument, e.g. deductive proof, hypothetical reasoning, reductio ad absurdum.
	D Dissecting an argument	Extracting and separating the relevant material from the less relevant (e.g. rhetoric, background). Identifying the key claims which might form parts of the argument.
	E Categorising the component parts of an argument and identifying its structure.	Recognising the parts of an argument and the function they play. E.g. evidence, examples, reasons *While "dissecting an argument" and "categorising component parts" often co-occur and work together iteratively, they are separate subskills*.
	F Identifying unstated assumptions	Looking for things (e.g. facts, beliefs, principles) which are essential to the argument but have not been explicitly presented.
	G Clarifying meaning	Detecting, avoiding and removing ambiguity for the purposes of reasoning soundly or judging the soundness of reasoning. Removing confusion over the meanings of words, phrases or expression of ideas that might alter the thrust or efficacy of the argument.
2 Evaluation	A Judging relevance	This process is more than simply judging relevant versus irrelevant. It entails judging the *degree* of relevance of a claim or piece of evidence to a particular interpretation or conclusion.
	B Judging sufficiency	Determining whether there is enough evidence to support a conclusion. Recognising the difference between necessary and sufficient conditions.
	C Judging significance	This entails judging the degree of importance of evidence in relation to conclusions and arguments.
	D Assessing credibility	Assessing the credibility of sources of evidence in relation to such criteria as expertise, corroboration or conflict, reputation, bias, factors that might interfere with accuracy of observation, judgement or reporting.
	E Assessing plausibility	In relation to claims, assessing the likelihood that a claim could be true, i.e "Is this the sort of thing which is likely to happen?" In relation to explanations, assessing the likelihood that the explanation given is the correct one (e.g. by considering alternative explanations).*This can often play an important role in assessing arguments*.
	F Assessing analogies	Judging whether two things being compared are sufficiently alike for the comparison to be useful (i.e. in clarifying and strengthening an argument).
	G Detecting errors in reasoning	Detecting errors in reasoning includes flaws in arguments, some common fallacies, incorrect inferences/deductions from information contained in a variety of sources (e.g. verbal, numerical, pictorial, graphical), as well as unfair manoeuvres such as irrelevant appeals e.g. to popularity.
	H Assessing the soundness of reasoning within an argument	Making an overall judgement as to how well the conclusion has been supported or justified by the argument as a whole. This will include considering the truth or plausibility of any of the individual claims or reasons, as well as the validity of reasoning (the degree to which the reasons support the conclusion.) The manner of assessment should be appropriate to the type of argument being assessed, e.g. deductive proof, causal reasoning, attempting to prove beyond reasonable doubt, attempting to establish likelihood based on balance of evidence.
	I Considering the impact of further evidence upon an argument	Judging the extent to which further evidence strengthens or weakens an argument. It may challenge, support, complement or conflict with evidence, reasons or unstated assumptions.
3 Inference	A Considering the implications of claims, points of view, principles, hypotheses and suppositions.	This requires looking at the wider implications of the components of the argument, including its overall conclusion. This will include checking for consistency and corroboration between the claims within an argument. Principles may be ethical principles.
	B Drawing appropriate conclusions	This involves ensuring the conclusion one draws is justified.
4 Synthesis/ construction	A Selecting material relevant to an argument	Gathering and collating appropriate and sufficient evidence.
	B Constructing a coherent & relevant argument or counter-argument.	Using one's knowledge of argument structure to construct one's own argument.
	C Taking arguments further	Extending an existing argument. Constructing new lines of reasoning which advance the argument.
	D Forming well-reasoned judgements[5]	Arriving at carefully considered and more accurate judgements in situations where there is insufficient evidence to allow certainty. (This involves applying all the relevant critical thinking skills)
	E Responding to dilemmas	This skill is applied in a situation where some action has to be taken in response to a problem, but any action taken will have undesirable consequences. It involves recognition of the consequences of competing courses of action, and an attempt to judge between them.
	F Making and justifying rational decisions	Deciding upon the best course of action once a conclusion has been drawn having applied the relevant Critical Thinking skills.
5 Self-reflection and self-correction	A Questioning one's own pre-conceptions	Gaining awareness of, examining and evaluating one's own pre-conceptions and being prepared to set them aside.
	B Careful and persistent evaluation of one's own reasoning.	Applying all of the above to oneself, with the aim of greater accuracy in one's own reasoning.

5 Judgement is wider than conclusion – it can mean a response, a decision.

Open-mindedness is an important aspect of CT. Being able to set aside one's own views is a pre-requisite for a fair examination of another's argument. Furthermore, open-mindedness allows a person to acknowledge that their own views may be unsupported or even wrong. Critical Thinking involves a fair assessment of evidence, rather than seeking to support or confirm one's own views.

The definition indicates that CT is a set of skills which one applies not only to other people's reasoning, but also to one's own. Being rational requires analysis, evaluation and elucidation of one's own thinking, with the aim of greater accuracy in one's own reasoning.

Other findings and observations

Mapping of Cambridge Assessment Critical Thinking qualifications and tests

There is only room here for an overview of the mapping findings. In brief, there were, as one might expect, differences in the combinations of sub-skills tested by the various tests, with only one sub-skill common to all, namely 'identifying conclusions'. There was very high congruence between any particular specification and its associated question papers. In just one or two cases, it was judged that some sub-skills were either evidently or implicitly sampled in the question papers or were apparent in the scripts, though not explicit in the specification. It was found that all Critical Thinking products were either substantially or entirely within the definition and taxonomy. Where specifications included sub-skills which were considered not to be Critical Thinking, this was usually attributable to intervention from external agencies.

Skills and Processes which are either on the fringes or more clearly outside the construct of Critical Thinking

Part of understanding what Critical Thinking *is* can be informed by understanding what Critical Thinking *is not*: identifying skills which are frequently confused with Critical Thinking, which lie close to the outer fringes, or may often occur concurrently with genuine Critical Thinking processes. Not all 'higher order thinking' is Critical Thinking.

1. **Reading comprehension**. Whilst reading comprehension is an underlying skill, it is distinct from Critical Thinking. Reading comprehension only asks what is in a passage and may be demonstrated through rephrasing, summarising or précis-ing. Reading comprehension does not, in itself, involve analysing or evaluating. At its closest to Critical Thinking, it involves clarifying the meaning of words or identifying the purpose.
2. **Problem solving**. This uses many reasoning skills and processes which are a facsimile of those in the Critical Thinking taxonomy. The main difference is that the solution to a problem (generally spatial and/or numerical) replaces the argument. Note that here a solution is defined as series of processes leading to the correct answer, and the 'answer' is analogous to a conclusion.The techniques for arriving at a correct solution in problem solving are in many cases different – e.g. trial and error and insight are much more important in problem solving than in Critical Thinking.
3. **Creativity**. An element of creative or imaginative thinking can sometimes be useful in assessing arguments and explanations (thinking up pieces of further evidence or alternative explanations which might undermine the reasoning) and in constructing one's own arguments or taking arguments further. Creativity is not an end in itself and nor is it an essential skill for Critical Thinking. For this reason, it is not contained within the taxonomy.
4. **Sampling issues in evidence**. Size of sample, representativeness, generalisability, understanding the role of a control group – this is all useful knowledge of experimental methods in social science, but in itself is not Critical Thinking. However, such knowledge can be useful to assess credibility and inferences from evidence (e.g. to help identify sweeping generalisations).
5. **Ethical content**, e.g. knowing the names and details of ethical theories, is not part of Critical Thinking. Knowledge of ethical principles, e.g. utililitarianism[6] and deontological theories[7], are on the fringes. Applying such principles and theories to a particular dilemma, however, does involve Critical Thinking.
6. **Syllogism**. This is on the fringes of Critical Thinking. Syllogistic arguments are rarely everyday arguments and, as such, the panel viewed syllogism as an irrelevant technicality for Critical Thinking.

It is hoped that this definition and taxonomy will provide a shared and common understanding of the construct of Critical Thinking. It provides a focus and a fixed reference point for future specification and assessment materials development work. Furthermore, it is hoped this definition and taxonomy will be valuable to teachers and students of Critical Thinking in providing clarity.

References

Black, B. (2007). Critical Thinking – a tangible construct? *Research Matters: A Cambridge Assessment Publication* **2**, 2–4.

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

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

Gill, T. & Black, B. (*in prep*). *Do candidates who have taken Critical Thinking AS level perform better in their A levels in other subjects?*

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

6 The doctrine that the greatest happiness of the greatest number should be the aim of social and political institutions.

7 Ethical theory concerned with rights and duties.

The future of assessment – the next 150 years?

Tim Oates Group Director, Assessment Research and Development

Parts of this article originally appeared in the Spring 2005 bulletin from the Tomorrow Project: 'Shaping the future? Or going with the flow?' They appear here reprinted with the kind permission of the project.

And in today already walks tomorrow Samuel Taylor Coleridge

Prediction is very difficult, particularly if it's about the future Niels Bohr

The paradox is that both of these quotes tap into truths about predicting the future shape of systems. What I will do in this article is look at trends and tendencies in the development of assessment, but also try to offer some theoretical perspectives on why developments take the shape that they do. Bohr is particularly interesting. With startling brevity, he introduces the idea that prediction in natural science is one thing, and in social science, something very different. Assessment systems are lodged in complex, highly interrelated social, political and economic systems. I will initially focus on this issue of what kind of science we can use to predict the future.

What most determines the shape of the future – the sum of individual actions? Ineluctable historical forces? The decisions of a powerful few? Attribution theory has been shown to be a powerful means of exploring why some people make greater progression in life and impact on events than others (Bem, 1972; Lepper *et al*., 1973; Miller *et al*. 1975). Some people feel carried along on a tide of events outside their control, whilst others feel as if they have personal agency – what they do has an effect and they can use this to enhance their world. These different groups attribute the cause of changes in circumstances which affect them to very different things. John Bynner's work at the Centre for Longitudinal Studies, on data from the people in the 1958 and 1970 cohort surveys, has allowed him to develop an insightful notion of 'personal capital' – personal resources upon which people can call to run their lives (Lambe, 2006; Schuller *et al*., 2004). Fundamental to this are feelings of personal power (or powerlessness). The people who display feelings of powerlessness tend to be those with worse outcomes in their lives – encompassing health, education, and social circumstances. His work shows that over time this makes a very substantial difference.

So, a notion that you are being carried along by externally-controlled events is bad for you (and your family). Alongside this, it is interesting to consider how people think of the way that economies, society and history develop. Phrases such as '… the natural operation of competition…', '… the tide of events…', '… the evolution of markets…' and similar crop up time after time in the media. They reinforce the idea of natural processes unfolding through their own unalterable dynamic. And the scale and subtlety of social and economic changes further promote these ideas of events and processes beyond human control – a shift in international markets that brings sudden unemployment to groups of workers and devastates specific communities; subtle changes in family structure brought about by both partners working full-time to sustain family income. Such changes seem far more related to 'natural social and economic evolution' than the results of specific human actions. It is a feeling which is compounded by our wish to attribute responsibility to someone, somewhere (Heider, 1944). This is further reinforced by the difficulty of changing the performance of important social institutions which affect our lives, such as education and health. They are juggernaut in size and structure – substantial investment and policies of direct intervention and change take so much time to bite and take so long to show results to increasingly impatient administrations.

But Realist social theory has re-cast the way we think about the impact of human action on the shape of social systems (Bhaskar, 1975; 1979). Social theories are a part of the social world – they affect the way the social world operates. Roy Bhaskar gives us an excellent example of this important perspective on social theory: the one pound coin. It's a round piece of metal which costs a great deal less than one pound. But it's worth one pound. Why? Because a group of people share a common belief that it's worth a pound. And I'm not knocking this. It is really useful that these shared beliefs operate in the social world. It enables the whole banking system, indeed the whole economy, to work. It shows us that beliefs play an important role in the operation of important social systems.

But while social theory and social research can be very good at explaining things – why certain social groups behave in certain ways – it is also notorious for its lack of precision in predicting events. Natural science is just great at predicting things – like the temperature at which water will boil when I take it up to 6000 metres, or the size of copper wire I will need to safely run a big piece of industrial kit. Frank Achtenhagen has outlined a powerful model of 'planned failure' in social policy (Achtenhagen, 1994). If you fail to adequately understand the nature of the problem you are tackling, you formulate policy which half-engages with the problem, but at the same time putting the policy in place changes the nature of the system you are dealing with, giving you a whole new set of problems which you no longer understand at all. This is the cause of the increasingly-mentioned 'unintended consequences' of policy.

This makes predicting the future a very difficult activity, since the future is partly composed of things which were intended and partly of unintended consequences, and is shaped by the shifting beliefs of people as well as objective forces. Some of these objective forces stem from factors such as limitations on natural resources, others from the impact of policy and action. Runs on banks are fascinating examples of the interplay of subjective and objective forces in social systems. They can be created by crises of confidence – confidence being a subjective human state over which people have individual control – but once people begin to act, based on that personal belief, the crisis becomes an all-too-tangible set of objective forces. They have the economic force and effect of a derailed express train, and appear as something over which individuals can effect little control.

With this as theoretical background I outline some key trends and developments in assessment. I do not advance them as 'the future', but as things which are most likely to feature in the set of factors which shape the future.

There is no shortage of analyses of the inertia in big public systems – interesting analyses of the attempts to reform pension systems, health systems and so on (Bramson and Buss, 2002; Donelan *et al*., 1999; Attwood *et al*., 2003). But the metaphor of 'inertia' does not do justice to the detail of the processes of reform. One new metaphor is needed to describe some of the efforts at change – something which captures a sense of the impetus required to escape the gravitational pull of existing arrangements. This metaphor may be of interest: you can launch a projectile into space on its way to new planets, but if it has inadequate energy, it will fall back to earth, and you end up near where you started, albeit at great expense and with quite a lot of wreckage. This captures the process which currently seems to be occurring in respect of national testing in England: new developments seem to lack the escape velocity to ensure that their purpose, form and operation are genuinely progressive. Innovations seem to be dragged back, by the pull of existing culture, opinion and processes, to a position where they mimic existing arrangements.

The new Single Level Tests were launched by their civil servant authors, in early 2008, as a radical development of national test arrangements (National Assessment Authority, 2008a). Responses to the consultation which followed the launch of the pilot for the tests suggested that the whole model was insufficiently distinctive from current arrangements, and that a range of fundamental measurement issues would prove troublesome in the piloting and operation of the tests. In the first administration of the tests (December 2008), many of these problems were indeed realised. Announcements have now been made (March 2008) regarding a shift in emphasis from using the tests to confirm that learners are 'secure' in a national curriculum level to 'threshold' performance in a level – that is, back to the current focus; and to explore the option of tests covering more than one level (BBC News online, 2008). If these changes are implemented, the supposed radical features of the new arrangements are to be diluted, and the testing arrangements will be far closer to simply providing two sessions, per year, of the existing test model. This brings the risk of testing further dominating the school curriculum (Mansell, 2007) – hardly the intended effect of the original innovation.

This tendency of initiatives to have inadequate 'escape velocity' has been evident in a series of major revisions to the education and training system. It has been particularly evident in vocational education and training. GNVQs are a prime example. Originally conceived with a radical project-based assessment model, GNVQs were constantly modified over a ten year period, each modification bringing the qualification closer and closer to existing assessment approaches in 16–19 general education. This was in part due to an attempt to increase 'parity of esteem' with academic qualifications, but also the result of a power struggle 'for the heart of the qualification' amongst Government agencies. By 2000, as GNVQs became Advanced Vocational Certificates of Education, the qualification had lost many of the features which were associated with learning programmes attractive to the original target group. The qualification had been dragged back to conformance with previous arrangements, no longer fulfilling the role and position which it had been designed for (Oates, 2008). This reduced significantly the range of vocational qualifications capable of being delivered in full time education.

Sometimes the 'pull of gravity' comes not from culture, or the predilections of policy makers, but from deeper structural factors. Although the picture is mixed in terms of quality and patterns of participation, Modern Apprenticeships at level 3 can broadly be considered a success – they are providing a well-grounded practical route to technician level employment. But the numbers of 16–19 years olds participating are startlingly low compared with other European countries which have an apprenticeship route. Total numbers on English apprenticeships at all levels, not just level 3, amount to barely 6% of the cohort, compared to 60% of the cohort in Germany. The causes of this are various, but derive mainly from the state of the labour market. With very low differentials between pay rates during training and pay rates for experienced workers, with training being viewed by hard-pressed employers as a short-term inefficiency, with licence to practice far less established in the UK labour market, and with wage flexibility a cornerstone of increasing employment rates and moving people from welfare to work, the structural conditions and incentive patterns simply militate against mass participation in apprenticeship. Under these conditions, you can try to make the form and content of the learning programmes and qualifications as attractive as possible, but participation simply is not going to undergo any seismic shift.

But whilst innovation is frequently dragged backwards by these processes, there are other societal, economic and technical developments which create constant pressure for change.

First, the explosion in information. The tendencies regarding blurring boundaries between 'private' and 'public' data are clear. In commerce, the patterns of data we leave behind us whilst purchasing goods and services are feeding huge systems of supply management and 'tailored' marketing – the latter presenting loops of feedback which determine in part how we see opportunity and how the commercial world is presented to us. 'Preferences' are recorded when we visit websites…personal profiles of 'you might like this…' built up and played back to us. The formative and summative assessment systems in place and under development fit into this pattern – increasingly fine-grained detail on individual performance, available not only to the learner, but also to teachers and managers of institutions, but also – of course with appropriate safeguards – to the state and its institutions.

In university admissions, in formative assessment in compulsory schooling, in all phases of education and training, there is increasing interest in the detail of performance – unit scores in A and AS examinations, attainment against the individual statements in the National Curriculum, profile components.

The problem here is that we can certainly generate this fine-grained information and we can develop increasingly sophisticated systems to store and display it. Some see the assessment future as being dominated by huge integrated school and college systems which simultaneously hold attendance records, personal data, all school management data (pay, room bookings), learning materials, summative data on individual attainment, formative assessment data, and so on. Apart from the vulnerability and dependency which such systems might stimulate, a key question for assessment is: are we matching our development of such systems with processes by which we can make valid inferences on the basis of these data? Our work with schools on formative assessment tools suggests that teachers do not yet have the skills or techniques to handle these complex arrays of data, and are not yet able to use the data as a basis for differentiated, 'personalised' learning to any great extent.

Richard Kimbell (2007) of Goldsmiths' College, working on the

innovative e-scape assessment project usefully reminds us: '…just because technology allows us to do exciting new things, it doesn't mean that we should do all of them'.

At national policy level, the availability of data on each and every child has led to increasing interest in accountability systems, the data being used as a system management tool within public policy. Many nations considering the future of their assessment arrangements are interested not only in using assessment for school and system monitoring and performance management, but also in international benchmarking – most notably to PISA, PIRLS and TIMSS. This marks a trend of assessment being the hub of control and comparison, as well as supporting more traditional functions associated with learning and progression. This is a heavy weight to carry.

What of the developments 'internal' to assessment? There are interesting things afoot.

The 'empty promise' of adaptive testing?

There was a huge flurry of interest in computer-based adaptive testing in the late 1990s, which waned with the publication of Wainer and Eignor's seminal 2000 review paper (Wainor and Eignor, 2000; Kreitzberg *et al*., 1997). Having expected much from tests which adapt to the performance level of candidates, thus promising greater reliability, reduced test length and/or greater domain coverage, ETS found adaptive systems to be expensive and patterns of item use peculiarly limited within banks – with acute problems of overuse and overexposure of a limited set of items. Expensive, elaborate systems were abandoned and general enthusiasm diminished. The ill-fated on-line KS3 ICT test developed by QCA, funded by the then DfES, started with the intention of having an adaptive model at its heart, but this was quickly abandoned as the complexities hit the development team.

Other issues remain problematic in adaptive tests systems: bank security; comparability problems associated with the facility of a test not being a simple sum of the facility of its items; comparability problems associated with each candidate potentially taking a unique or near-unique combination of items (*op cit*). But small groups of developers have quietly worked away at the provision of working systems – the ESOL group at Cambridge Assessment, Peter Tymms and colleagues at the CEM Centre at Durham, and effective operational systems with robust measurement characteristics are beginning to emerge. Adaptivity may be maturing and emerging as an interesting solution to some of the more enduring problems of mass assessment: the problems of designing single assessments which are accessible to large populations of learners of widely varying levels and patterns of achievement, problems of tiered papers, with their well-known, vicious problems of ill-managed entry/access strategies and equity issues associated with floor and ceiling effects.

On-demand 'test when ready' approaches

'Testing when ready'; 'stage not age', driven by concepts of 'personalised learning' have surfaced as powerful guiding principles for public policy on assessment (BBC News online, 2007). I discuss elsewhere the problems that this may be only superficial rhetoric, with the 'gravity' of existing models and mechanisms pulling innovations back to older, existing models and modes of operation. But these new concepts are nonetheless proving powerful shapers of policy discourse. The new Single Level Tests (SLTs), under pilot in 10 LEAs, are intended to deliver, through six-monthly test opportunities, 'testing when ready' and 'stage not age' assessment. With six-monthly test occasions, candidature in national testing will remain very substantial – with many potentially taking tests more frequently. But even under these conditions, the nature of the entry arrangements pose potential threats to statistically-based standards-maintenance processes. Relatively stable, high population entry is essential to the kind of standards-maintenance processes which are currently used in most educational tests and examinations in England. The potential for only ever having low numbers taking the tests at any given moment (in a fully-blown on-demand system) affects not only the award process but also the ability to see quickly through statistical monitoring any peculiar patterns pointing to defects in the tests/test items. Only having 'when ready' candidates will affect attempts to maintain standards over time, where current fixed test sessions include a mix of 'ready' and 'less ready' candidates. In one legitimate interpretation of 'when ready' testing, an assumption can be made that pass rates should be close to 100% – certainly, the issues of who decides when a person is 'ready', and what the operational definition of 'when ready' actually is, remain problematic.

The drive to 'authentic' tasks

Advocates of ICT-based assessment frequently cite the possibility of setting more complex (aka 'rich', 'dense', 'textured') assessment tasks which assess 'higher order' skills (National Assessment Authority, 2004). This is assumed to be an unmitigated benefit, but the scoring processes, equity issues (in particular the complexities of the tasks and the need for candidates to be clear about what they need to do to succeed in the task), and what constructs are actually being assessed remain highly problematic. An under-recognised issue is that new forms of test may invoke different forms of cognitive engagement. This is illustrated by airline pilot assessment using simulators – you actually want the pilots to believe fully in the test that they are flying – that is, to have full cognitive engagement and no longer be conscious that they are being tested. Should this be emulated, indeed be a goal, in educational testing? There is clear evidence that maintaining awareness of what the test is actually asking for (e.g. seeing past the 'scaffolding') can elevate test performance and can enhance learning. How will tests which emulate the 'simulation' paradigm affect equity (access) for different groups? There can be no simple assumptions that high authenticity, complex tasks should be an ideal in educational testing.

The technological transformation of assessment

Meanwhile, the technological transformation of assessment continues apace, with few commentators doing anything other than picking up on one or two of the full set of ways in which assessment is indeed being transformed:

- Production of assessments (item banking, 'paperless' preparation of 'traditional' exam papers which are then sent direct to printers and then despatched to schools, archiving of materials for reference in comparability studies and standard-setting).

– Provision of on-demand testing, of rapid feedback, and formative assessment.
– Automation of marking of both objective and open response items (automated systems, including those using artificial intelligence – something I deal with below).
– Allocating learners to 'levels' (tiered exam papers replaced by adaptive on-screen tests).
– New ways of presenting questions on screen (development of new types of questions such as those showing rotation of three-dimensional objects, simulations, etc).
– Response by candidates (new types of responses to stimulus material, such as dragging and dropping material).
– Management of scripts (electronic script management – scripts are scanned in and can be sent to markers).
– Restructuring of marking activities (e.g. giving markers the same question from different candidates' papers rather than whole papers to mark).
– Management of results (electronic result management, e.g. texting results to candidates).
– Operation of quality assurance models (e.g. real-time monitoring of markers as they mark on-screen and intervening if problems occur).
– Integration of assessment, learning and MIS information (big school-wide systems).
– Evaluation and research (using scanned scripts and results to run simulations, in order to explore the impact of new assessment processes, but without prejudicing real candidates' chances; integrating assessment data with other data on candidates, such as social background).

Much of the seemingly parochial 'backroom' work on electronic management of question-paper construction, electronic management of scripts (and thus the possibility of new quality assurance processes for marking) is having a huge impact on qualifications. The development of item-level analysis holds huge promise for enhanced quality assurance processes and for research. But the detail of systems matter – not being able go back through marking is a serious weakness in some of the on-screen marking systems; using some forms of scanning prevents markers' annotations from being recorded; … and many of these systems are not so much stable applications as enormous, continuing development projects. But the prize here is almost certainly not reduced cost, but increases in quality and service.

Finally, a few 'emerging issues':

The rise of 'outcomes-based' qualifications in vocational education and training – revised paradigms?

There is a strong international trend towards outcomes-based qualifications (independent of the mode, duration or location of learning) – it is an approach that is reinforced by the commitments intrinsic in the European Qualifications Framework (Oates, 2004; European Commission, 2008). This has the effect of placing high demands on assessment including mastery approaches, high coverage of all necessary skills and knowledge. In addition, it is clear that the concepts of competence embedded in these approaches are crude, and underestimate the importance of vital processes of 'professional formation'. If 'competence-based' models begin to intrude on educational assessment, one of the most important areas to watch will be 'mastery' versus 'compensation' – with mastery tending to demand performance in all elements, thus pointing towards a series of low hurdles rather than items with strongly contrasting facilities/demands. This would represent a fundamental switch in measurement paradigm. Interestingly, this is an important problem facing the policy-makers and developers involved in the Diploma initiative.

Increased enthusiasm for teacher assessment – evidence of benefit in the English setting?

The reviews by Daugherty (Wales) (2004) and Tomlinson (England) (2004) asserted a need to increase the role of teacher assessment in national systems. Neither review presented evidence that teacher assessment can operate in such a way as to deliver stable assessment outcomes in a context of high stakes accountability arrangements. Indeed Sweden offers evidence to the contrary, with acute 'grade inflation' accompanying the introduction of national accountability systems in a system relying heavily on teacher assessment. The principal example of teacher assessment advocated by policy makers etc (Queensland) has not yet integrated accountability arrangements, nor has it generated data on standard reliability measures etc. Classification error is thus difficult to establish – a crucial problem. While the enhancement of learning remains an apparent benefit of such arrangements, the introduction of teacher assessment into a context overdetermined by high stakes accountability arrangements remains highly problematic. What is needed is well-designed research on the technical characteristics of teacher assessment under different system conditions. Without this, a drive towards teacher assessment could well be a leap of faith, in the dark. This carries worrying ethical implications.

Tiering – sufficiently equitable?

Linked to the above, tiering is designed to address clear problems of designing papers which are pitched at the right level for 'bands' of learners (with the specific intention of allowing learners to best demonstrate what they know and can do), but with each specific model for tiering exhibiting undesirable artefacts and deficits. Will tiering continue to be considered by assessment specialists, educationalists, parents and learners as being sufficiently equitable?

Levels (and grades) – are they sustainable?

While the national curriculum legislation requires reporting of children's level of attainment to parents (National Assessment Authority, 2008b), the diagnostic and informative capacity of 'levels' remains under-researched. What do parents make of 'your child is at level 4'? Does it help them direct their support at home in the best way possible (evidence of continued social inequalities in educational outcomes suggest it does not help all families equally)? Levels are blunt; a reduction in diagnostic content in contrast to the scores which make them up – as

Ian Schagen (2003) of NFER has stated in a number of contexts, we spend half our time working scores up into levels and then the rest of the time breaking them back down again to make educational sense of them. Levels introduce a discontinuous scale, with all the attendant problems of two pupils immediately either side of a level boundary being more alike then two pupils at extreme ends of the same level. Misclassification at a key point in a person's educational progression can lead to radically different (inappropriate) educational treatment. The artefacts, identified by QCA's own researchers, around the level thresholds are highly problematic. But many grading systems exhibit similar problems. Both levels and grades may fall foul of increasing public concern for equitable treatment in both access to learning and in educational measurement.

Attacks on the possibility of maintenance of standards over time – and intolerance of measurement error

The gap between public understanding of assessment and expectations of technical rigour remains wide (Wood, 1993; Newton, 2005). There is increasing commitment of assessment specialists and managers to enhance public understanding, with the pressure this brings for more realistic expectations regarding the difficulty – and indeed the sense – of maintaining standards over anything but short time frames. Concern over maintaining standards may be increasingly replaced by concerns to ensure that qualifications are fit for purpose in respect of ever-changing societal, labour market and economic requirements.

Whilst this article may not offer the apparent certainties peddled by futurologists (warning: believe them and that might make it true), it tries to map out some of the trends and tendencies which are playing a part in shaping the future. Perhaps the most important message from this analysis is that our intentions DO matter – the values which we hold will shape events and systems. Clarity of purpose and a firm accountability to learners would seem to be a vital bedrock under the shifting sands of public assessment systems.

References

Achtenhagen, F. (1994). *How should research on vocational and professional education react to new challenges in life and in the worksite*. Paper presented to International Research Network on Training and Development (IRNETD) Conference, Milan, March 1994.

Attwood, M., Pedlar, M., Pritchard, S., & Wilkinson, D. (2003). *Leading change – a guide to whole system working*. Bristol: Polity Press.

BBC News online (2007). *'Testing when ready' gets going* http://news.bbc.co.uk/1/hi/education/7137149.stm Accessed 11 04 08.

BBC News online. (2008). *Pilot progress tests made easier*. http://news.bbc.co.uk/1/hi/education/7246871.stm Accessed 11 04 08.

Bem, D. (1972). Self-perception theory. In: L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 6. New York: Academic Press.

Bhaskar, R. (1975). *A realist theory of science*. Falmer: Harvester.

Bhaskar, R. (1979). *The possibility of naturalism: a philosophical critique of the contemporary human sciences*. London: Harvester.

Bramson, R. & Buss, T. (2002). Methods for Whole System Change in Public Organizations and Communities: An Overview of the Issues. *Public Organisation Review*, **2**, 3, 211–221.

Daugherty, R. (2004). *Learning pathways through statutory assessment: key stages 2 and 3. Interim report of the Daugherty Assessment Review Group*. Daugherty Assessment Review Group, Cardiff.

Donelan, K., Blendon, R.J., Schoen, C,, Davis, K., & Binns, K. (1999). The cost of health system change: public discontent in five nations. Harvard Opinion Research Program, Harvard School of Public Health, Boston, USA. *Health Affairs*, **18**, 3, 206–16.

European Commission. (2008). http://www.europe.org.uk/news/view/-/id/218/Accessed 11 04 08.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, **51**, 358–374.

Kimbell, R. (2007). *Technology and the assessment of creative performance*. Technology Education Research Unit, Goldsmiths College, London. Keynote presentation at Cambridge Assessment Conference, 15th Oct 2007.

Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1997). *Computerized Adaptive Testing: The Concept and Its Potentials*. Princeton: Educational Testing Services.

Lambe, B. (2006). *Conceptualising and measuring agency using the British Household Panel Survey data*. BERA annual conference 2006, University of Warwick.

Lepper, M., Greene, D., & Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, **28**, 129–137.

Mansell, W. (2007). *Education by numbers: the tyranny of testing*. London: Politico's Publishing.

Miller, R., Brickman, P., & Bolen, D. (1975). Attribution versus persuasion as a means of modifying behaviour. *Journal of Personality and Social Psychology*, **31**, 430–441.

National Assessment Authority. (2004) http://www.naa.org.uk/naaks3/documents/2004_KS3_ICT_report.pdf Accessed 11 04 08.

National Assessment Authority. (2008a). http://www.naa.org.uk/single_level_tests/ Accessed 11 04 08.

National Assessment Authority. (2008b). http://www.qca.org.uk/eara/ Accessed 11 04 08.

Newton, P. (2005). The public understanding of measurement inaccuracy. *British Educational Research Association Journal*, **31**, 419–442.

Oates, T. (2004). The role of outcomes-based qualifications in the development of an effective vocational education and training (VET) system. *Policy Futures in Education*, ISSN 1478–2103, **2**, 1.

Oates, T. (2008). Going round in circles: temporal discontinuity as a gross impediment to effective innovation in education and training. *Cambridge Journal of Education*, **38**, 1.

Schagen, I. & Hutchison, D. (2003). Adding value in educational research – the marriage of data and analytical power. *British Educational Research Journal*, **29**, 749–765.

Schuller, T., Bynner, J. & Feinstein, L. (2004). *Capitals and capabilities*. Centre for Research on the Wider Benefits of Learning.

Tomlinson, M. (2004). *14–19 Qualifications and curriculum reform*. Nottingham: DfES.

Wainer, H. & Eignor, D. (2000). Caveats, pitfalls and unexpected caveats of implementing large-scale computerised testing. In H. Wainer (Ed.), *Computerized adaptive testing: a primer*. Chapter 10. 2nd edition. New Jersey: Lawrence Erlbaum Assoc Inc.

Wood, R. (1993). *Assessment and testing*. Cambridge: University of Cambridge Local Examinations Syndicate.

Cambridge Assessment marks 150 years of exams

Jennifer Roberts Public Affairs

For 150 years Cambridge Assessment has been at the forefront of enhancing education through assessment – and our success owes much to the contribution of our examiners and partners around the world.

There have been many changes to the education system over the years but the ethos that sparked the creation of the University of Cambridge Local Examination Syndicate (UCLES) – the former name of Cambridge Assessment – still drives our work today. We continue to strive for the ongoing improvement to assessment systems and methodologies used around the world to ensure learners access the benefits of their education.

The first UCLES exams took place in 1858 in eight regions in the UK. Students sat examinations in English Language and Literature, History, Geography, Geology, Greek, Latin, French, German, Physical Sciences, Zoology, Chemistry, Arithmetic, Mathematics, Drawing, Music and Religious Knowledge (unless parents objected). Today, Cambridge Assessment delivers assessments in 150 countries.

A series of special events is taking place throughout 2008 to mark our 150th anniversary:

- To show just how far the exam process has come, we visited some of the schools that sat the first 'Cambridge' school exams. Today's pupils experienced an 1858 style lesson and attempted questions from the original UCLES exam papers. Pupils commented that they found their lesson an educational and eye-opening experience.
- On 11 February, 150 years to the day that the Syndicate was officially established, a commemorative book, *Examining the World*, a collection of essays on the development and immense changes that have taken place in the world of exams in the UK and worldwide, was launched. The book is published by Cambridge University Press and available from www.cambridge.org
- An online version of our exhibition featuring more than 40 reproduced documents and photographs from our archives was unveiled on our website at www.cambridgeassessment.org.uk. The exhibition includes bribery letters, 150-year-old examiner reports, eye-witness accounts of hardship during the First and Second World Wars and past exam questions.
- In April, the Association of Language Testers in Europe (ALTE) 3rd International Conference was hosted by Cambridge ESOL in Cambridge. The conference theme, *The Social and Educational Impact of Language Assessment*, formed a bridge between the world of language assessment and educational, social, cultural and economic environments and contexts.

Forthcoming event: 34th International Association for Educational Assessment (IAEA) Annual Conference

From 7–12 September 2008, Cambridge Assessment will host the 34th IAEA Annual Conference in Cambridge. The IAEA Annual Conference is a major event in assessment, bringing together leading assessment and education experts and providers of examinations from across the world.

The conference theme is *Re-interpreting Assessment: Society, Measurement and Meaning*. The keynote speakers will be Professor Robert J Mislevy, University of Maryland, and Professor Dylan Wiliam, Institute of Education – University of London.

Registration will remain open until 11 July 2008 and further details can be found at www.iaea2008.cambridgeassessment.org.uk

Research News

Conferences and seminars

3rd International Rasch Measurement Conference

In January Tom Bramley attended the 3rd International Rasch Measurement Conference in Perth, Western Australia and presented a paper entitled 'Maintaining performance standards using expert judgement: a rank-ordering method'. Topics discussed at the conference included Rasch model applications in education, computer adaptive testing, developments in Rasch modelling and item banking.

American Educational Research Association (AERA)

Beth Black attended the AERA conference in New York in March and presented work on maintaining performance standards using expert judgement.

Association of Language Testers Conference (ALTE)

The ALTE 3rd International Conference, hosted by University of Cambridge ESOL Examinations, took place in Cambridge in April. The theme of the conference was *The social and educational impact of language assessment*, and the aim was to form a bridge between the world of language assessment and educational, social and economic environments and contexts. Gill Elliott, Nat Johnson and Sylvia Green of the Research Division presented their paper on *'Aspects of Writing: Using an atomistic approach to evaluate qualities of features of writing*.

House of Commons Research Seminar

The third House of Commons Research Seminar, chaired by Barry Sheerman MP, Chair of the Children, Schools and Families Select Committee, was held on 24th January. The theme was *What makes a good teacher? An overview of teaching effectiveness research*. There were presentations from Professor Patricia Broadfoot, Professor Mary James and Professor Debra Myhill followed by a plenary session. The aim of these seminars is to bring together members of the research, academic and education communities as well as policy makers and influencers.

New Horizons seminar

In April Tom Bramley, Beth Black and Tim Gill from the Research Division gave a seminar in Hughes Hall, Cambridge on *The rank-ordering method for maintaining standards by expert judgement*. The rank-ordering method is a new technique for equating the raw score scales on two tests by using expert judgement. The judgemental task involves experts making relative, holistic judgements about examples of student work from two tests and combining these into a single rank-order.

This seminar discussed the development of the rank-ordering method out of the Thurstone paired comparisons method used in UK comparability research, and showed how it has been applied in several standard-maintaining contexts. The seminar then discussed some theoretical and practical issues raised by the method, before finally considering how it could capitalise on new developments in on-screen marking.

British Educational Research Association Conference, 2008

The annual conference of the British Research Educational Conference (BERA) will be held at Heriot-Watt University, Edinburgh, from 3–6 September. Cambridge Assessment will again be providing sponsorship.

The keynote speakers will be Professor Lindsay Paterson from the University of Edinburgh, and Professor Gloria Ladson-Billings and Professor Ken Zeichner, both from the University of Wisconsin-Madison.

Professor Paterson's keynote address will deal with the implications of some recent research on social mobility and education in Scotland during the twentieth century, which he will use to seek to illuminate some wider questions about opportunity and democracy. Professor Gloria Ladson-Billings will explore issues of race and educational disparities and the need to look not only at the schools, but rather at the 'new' students who are attending schools and their part in the school landscape. Finally, Professor Zeichner will focus on the interaction – or lack of interaction – between the different worlds of practitioner research and academic research, focussing on the 'third space' theory where these worlds can be joined for mutual benefit.

In a departure from the traditional keynote address at conference, BERA 2008 will also feature an 'expert' panel that will present and answer questions around a theme of social justice.

Eleven researchers from the Research Division will be attending the conference and the following papers, covering a wide range of themes, will be presented:

- **Beth Black:** Investigating Non-Standard English in GCSE level students in England
- **Tom Bramley:** Mark scheme features associated with different levels of marker agreement
- **Victoria Crisp:** Judging the grade: An exploration of the judgement processes involved in A-level grading decisions
- **Victoria Crisp and Nadezda Novakovic:** Are all assessments equal? The comparability of demands of college-based assessments in vocationally-related qualifications
- **Gill Elliott:** Teaching Practical Cookery in UK schools
- **Gill Elliott and Nat Johnson:** Spelling mistakes at age 16. A detailed analysis of the nature of spelling errors encountered in a sample of GCSE English writing
- **Tim Gill and Tom Bramley:** How accurate are examiners' judgments of script quality in the absence of information about mark totals? An investigation of absolute and relative judgements in two A-level units
- **Jackie Greatorex and Rita Nadas:** Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions?
- **Nat Johnson:** An evaluation of the development of a multiple-item scale for assessment validation in relation to classroom practice
- **Nat Johnson and Gill Elliott:** An atomistic approach to comparability of student performance in mathematics
- **Irenka Suto and Rita Nádas :** Towards a new model of marking accuracy: An investigation of IGCSE biology
- **Carmen Vidal Rodeiro, Joanne Emery and John Bell:** Can emotional and social abilities predict differences in attainment at secondary school?

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

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

ISSN: 1755–6031