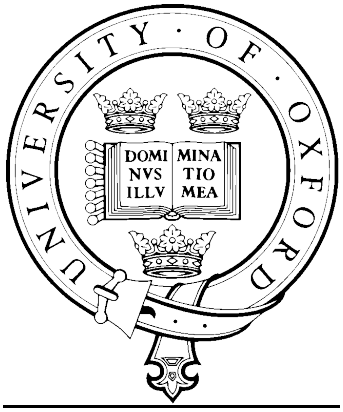

Marking time: the decision-making processes of examiners of History and English 'A' level

Victoria Faith Elliott

Exeter College

Supervised by Professor Ingrid Lunt and Professor Gordon Stanley

Oxford University Department of Education

Oxford University Centre for Educational Assessment

Abstract

In the UK examiners assign marks to A level examination scripts using extensive mark-schemes. Examiners work under strict time constraints, and must consider various sources, from script to mark-scheme to exemplar marked scripts. Essay subjects, such as History and English (two subjects which are associated with difficulty of marking), are likely to form a particular cognitive challenge for examiners, and their marking has not been extensively researched. Most examiners mark within accepted variations of reliability, as determined by Awarding Bodies' monitoring systems. The question is, then, how they make these decisions, given the amount of information and the limited time available. The training process which is intended to bring the examiners' decisions into line with the standard of the Principal Examiner also represents a lacuna in the literature. This study therefore sought to investigate examiners' decision-making processes and the process of the training meeting.

Five day-long standardising meetings (four examiners' meetings and one senior examiners' pre-standardising meeting), split between English and History, were recorded, transcribed and subjected to discourse analysis; three examiners, spread between the four units, provided additional Verbal Protocol Analysis data while undertaking live marking. A survey, which presented preliminary conclusions from that data and some extracts from examiners' discourse, was used to collect further data from a larger sample of History and English A level examiners. The data are considered in relation to the theory of heuristics, which has been used to consider examining at other levels, or in other subjects, and with other question types. The data are also considered in the light of other theories, including those of expertise and construct-referenced assessment.

The data demonstrate that decisions were not usually made in the rule-based way which is suggested as the ideal by the regulations, and which would be assumed from the mark-schemes and rubric issued by the exam boards. The mark-scheme did provide a guide to the foci which should, and can be seen to, attract examiners' attention. However, a great deal of 'professional judgement' was also exercised, and examiners used a number of informal heuristics, and made relative judgements to reach a mark; comparison is established as a major mechanism of their decision-making. These behaviours do not necessarily lead to bias, however, and many were actually suggested during the training process. Some were suggested consciously by senior examiners, but some appeared to be unconsciously modelled during the training meeting. The theory of heuristics is seen to be widely applicable to the data; the choice of material and training mitigated the potential bias which heuristics could cause. A wide range of cognitive processes are demonstrated in the data, which were used to varying degrees by different examiners, at different times within and between scripts.

Acknowledgments

I am very grateful to the Awarding Body who hosted my research. Every single examiner who volunteered to participate, whether as a recordee, a 'think aloud' participant or a questionnaire respondent: you are the people without whom research could not happen. Lupin, Caspar and Parrot, thank you particularly.

The most heartfelt thanks go to both my supervisors, Gordon Stanley and Ingrid Lunt. Gordon shaped this project into what it is, and networked tirelessly on my behalf. Ingrid not only supported me before there was anyone else in the department working on assessment, she also seamlessly transferred from joint to single supervisor when Gordon retired back to Australia; her minute attention to detail is vastly appreciated.

The support staff at OUDE are unfailingly kind and helpful. I particularly would like to thank Jo Hazell for her cups of tea and friendly chat across the OUCEA office. The library staff are all heroes; Kate and Shez have been especially helpful over the years.

Fellow students provide the kind of support that we all need; Beth Hore, with whom I exchanged despairing emails on the subject of literature reviews and conference papers, is always a little treasure. Lila McDowell kept me going in our 'virtual' group of two, with cross-Atlantic emails, a bed in NYC, cat-playing duties and late night meals. She also kindly read various things for me, prodded me when necessary and laughed at my poor theory-based jokes. Stephanie Berry's occasional cheerleading emails and comfort packages from Oxford, MS, kept me going. The ceramic frog telling me to 'hop on it' definitely contributed! Abby Loebenberg's philosophy that as a DPhil student you can really only forget about your thesis when engaged in an extremely physical activity in which things fly at your head was justification enough to play with OUWRFC. Abby, (BF)SJ and all the other forwards, Panthers and Blues, have been mainstays of my time at Oxford, and ensured my survival through this exercise in emotional tenacity. Back yourself.

I'd like to thank Sharon Greig and Gemma Hall for their happiness to act as guinea pigs for my research instruments. My sister Robynne kept me company in the last year, and my sister Emily provided secretarial services beyond compare. One helped me move into Oxford, and one helped me move out; you're both stars. The support of family and friends, too numerous to mention is what every DPhil student needs. I love you all.

I grew up thinking that a doctorate was what you did, because between the ages of four and seven, I saw my mother doing just that. Now I can appreciate the enormity of an achievement that included earning a living, looking after me *and* finishing in three years. So, just as hers was dedicated to me, this is for her: for Geraldine.

Contents

Abstract	**1**
Acknowledgments	**2**
Contents	**3**
List of abbreviations	**4**
Introduction	**5**
Chapter 1: Theoretical framework	**12**
Chapter 2: Examining the literature	**29**
Chapter 3: Methodology: How to look inside the black box	**52**
Chapter 4: Characterising the training meetings	**68**
Chapter 5: Analysis of scripts – foci of attention	**93**
Chapter 6: The role of the mark scheme	**116**
Chapter 7: Comparison: a principle of decision-making	**132**
Chapter 8: Heuristics – cognitive shortcuts for human judgement	**152**
Chapter 9: Questionnaire findings	**169**
Chapter 10: Conclusions: addressing the research questions	**189**
References	**203**
Appendix: Questionnaire and breakdown of statements by theme	**216**

List of Abbreviations

AO	Assessment Objective
AQA	Assessment and Qualifications Alliance
OfQual	Office of Qualifications and Examinations Regulation
QCA	Qualifications and Curriculum Authority
QWC	Quality of Written Communication
VOIP	Voice Over Internet Protocol
VPA	Verbal Protocol Analysis

Abbreviations for data attributions

ChaE	Chair of Examiners
CE	Chief Examiner
E1	English 1
E2	English 2
H1	History 1
H2	History 2
PE	Principal Examiner (PE1 & PE2)
SE	Senior Examiner

Introduction

Overview

Examining is widely assumed to be a rule-based, structured and highly rational activity, which supports the requirements of fair assessment. Such characteristics are prized by the regulator of examinations in the UK, OfQual, whose code of conduct prioritises 'consistency' and explicitly demands procedures which are in 'the interests of reliable marking and [which] reduce the scope for variability', providing examiners 'with detail that allows marking in a standardised manner' (OfQual, 2010: *passim*). However, it is clear that the length and complexity of the answers, particularly in the humanities, coupled with what represents sufficient 'detail' must challenge the underlying assumptions of the examining process.

Examinations for qualifications such as the A level and GCSE in the UK rely on a panel of external examiners, which requires that they hold a common understanding of standards and mark schemes. In the attempt to ensure this, the Awarding Bodies provide examiners with detailed mark schemes containing thorough criteria for the judging of quality. Examiners then undergo a carefully prepared training process which is designed, together with the mark scheme, to eliminate as much of the subjective element of human judgement from their marking as possible.

The normal procedure in Awarding Bodies is for examiners to be invited to a standardisation meeting, before which they are sent a copy of the exam paper and the standard rubric for marking the paper. For longer responses, or essay type questions, the rubric tends to be generic, remaining largely the same from year to year. At the standardisation meeting examiners are provided with scripts which have been graded and agreed by the principal examiner and other senior examiners, to serve as models. The purpose of the meeting is to 'catch the standard', that is, to ensure that all examiners are applying the generic rubric in the same way as the principal examiner for that paper. Examiners are expected to read the answers of students, which are often lengthy, and then grade them according to the provided criteria. The mark scheme for a single essay question is a minimum of an entire page of A4. Particularly in the case of essay questions the cognitive load for the individual examiner may be very high, exceeding what may be reasonably processed and used, if we invoke Simon's concept of 'bounded rationality' (1992).

Essay-based examinations have been widely recognised as problematic since the beginning of the twentieth century (Thompson and Bailes, 1926), in terms of setting,

answering and assessing. History and English in particular are subjects in which a high level of subjective judgement contributes to assessments of quality. However, awarding bodies consider their examiners to consistently achieve an acceptable level of reliability, according to the measures which OfQual (or previously, QCA) requires them to use. Examiners cannot be using the mandated procedure to reach their grades for essays in these subjects. Yet they are reaching the 'correct' decisions at a reasonable speed.

There is a clear need, therefore, for research which seeks to explore the decision-making processes of examiners, and seeks to consider them in the light of current theories of cognitive psychology. Milanovic & Saville emphasised the importance of understanding the decision-making behaviours of markers because 'lack of knowledge in this area makes it more difficult to train markers to make valid and reliable assessments' (1996:94).

It is important to note that it is not potential causes of bias that are of interest; much research has already been done on sources of unreliability in examiners' marking, such as influence caused by gender, candidate's name and the like. This study seeks to explore the mechanisms by which competent examiners are making their judgements on candidates' work, when those judgements fall within the range deemed acceptable by the awarding bodies.

The examination system

The General Certificate of Education Advanced or 'A' level is the qualification taken by most UK students at the age of 18. Each subject will typically consist of four modules, two examined internally and two externally, with one of each taken in each year of the sixth form. In subjects such as English and History an externally examined unit, with which this study is concerned, is usually composed of two essays, often chosen by the candidate from a range of potential questions. There are two examination sessions, one in each of January and June, with the larger pool of candidates sitting the examinations in June. Consequently the pool of examiners is also larger in June; those who are asked to mark in January are selected from the larger pool on the basis of their reliability.

After candidates have sat the examination, a copy of the question paper and the mark scheme are sent to examiners. Mark schemes require analytical rather than holistic marking, with the use of 'Assessment Objectives', which are subject-specific; any given unit will cover a sub-set of these, so that overall a qualification takes performance in all of them into consideration. The following is an example set of assessment objectives from an English Literature A level:

> **AO1** Articulate creative, informed and relevant responses to literary texts, using appropriate terminology and concepts, and coherent, accurate written expression
>
> **AO2** Demonstrate detailed critical understanding in analysing the ways in which structure, form and language shape meanings in literary texts
>
> **AO3** Explore connections and comparisons between different literary texts, informed by interpretations of other readers
>
> **AO4** Demonstrate understanding of the significance and influence of the contexts in which literary texts are written and received (AQA, 2007)

A mark scheme will provide separate criteria for the two or three assessment objectives for a question. Criteria are separated into a number of 'bands' or 'levels', within each of which there is a range of marks which may be awarded. Typically each band will have two or three bullet points within it, which demonstrate a progression of skill in a given area, through the change in adjectives or adverbs which describe the performance. A mark scheme for one question will usually cover a page or more of A4, with the assessment objectives appearing in parallel or sequence. The mark scheme is generic and remains the same from examination session to examination session. In addition, examiners may be provided with 'indicative material', usually another page of A4, which contains information that may be contained in the answer to a question; it is designed neither to be exhaustive nor exclusive. This material is dependent on the question which has been asked and is therefore different for each examination.

In this period between the examination and the standardisation meeting examiners are expected to access answers given by candidates through the online system, or to read physical scripts which have arrived at their house via post. They do not mark, but read enough scripts to familiarise themselves with the range of answers which may be given. By the time they arrive at the standardisation meeting, therefore, they are expected to be thoroughly familiar with both questions and mark schemes. The short period of time and the fact that most are also practising teachers, however, means that this expectation is not always met.

Meanwhile the most senior examiner on the paper, the Principal Examiner, who is ultimately responsible for setting the questions and whose judgement is the standard for the examiners to meet, gathers a range of scripts and makes judgements about what marks they deserve. He or she will consider as many examples as possible in the time, and attempt to

find a selection of answers which cover the range of attainment, across the range of questions. These scripts form the basis of the standardisation meeting. In examinations which have a large number of examiners, the Principal Examiner will hold a pre-standardisation meeting, to train the Team Leaders, or Senior Examiners, who will supervise and lead small teams of examiners. The following day a standardising or training meeting is held, in which the main body of markers are trained; in small examinations this is the only meeting which is held, with all examiners effectively acting as one team under the direct supervision of the Principal Examiner.

After the meeting examiners will go away and mark a 'sample', consisting of a small number of scripts drawn from their total allocation (a typical allocation will be composed of 120 to 250 scripts, to be marked over a short time period, which can be as little as eight days in the January session, or as much as three weeks, more usual in June). Their supervisor will check the decisions they have made about the mark which the scripts deserve, and give feedback about the decisions which examiners are making. If they are operating within acceptable limits, that is, if the standard of their judgement is close enough to the standard of the Principal Examiner, they are permitted to continue marking at will. If they are close, but not quite within acceptable limits, they will be given advice and asked for a second sample, after which a decision is made as to whether they are able to mark scripts. In some units, and if the marking period is long enough, another sample of marking is made towards the end of the marking period, to ensure the continued reliability of the examiner. At the end of the examination session examiners are issued with a grade for their marking; only those with A or B ratings are invited to return in future. Reliability in terms of UK examinations is a very particular measure of inter-rater reliability between a single examiner and the judgement of a senior examiner. Where there is a large number of examiners operating in teams, the team leaders undergo a similar process with the Principal Examiner monitoring their marking: the standard is passed down the chain of command.

Much marking is now done via an online electronic system; this means that senior examiners can and do check the marking of scripts at random, picking out scripts from the system to 'backread'. If examiners are considered to be not maintaining the required standard, then they are stopped from marking, either until they have been re-trained, or permanently.

After the marking has been completed, the Principal Examiner and some senior colleagues engage in a process known as 'grading', a further meeting at which the mark boundaries which differentiate between A and B grades and between E and U grades (the

pass/fail border) are set, using marked scripts from the examination session. Other boundaries are statistically determined (Crisp, 2010a).

The vast majority of scripts, therefore, are marked only by a single examiner: for this reason it is especially important for Awarding Bodies to ensure that all examiners are making judgements of a very similar standard, to provide a fair assessment.

A note on terminology

Principal Examiners are, as stated, the individuals who are responsible for the questions and the marking standard on a unit. They are answerable to the Chief Examiner, who has overall responsibility for a qualification. The individuals who lead groups of examiners, under the Principal Examiner, are known as Senior Examiners or Team Leaders. The phrase 'senior examiners' without capitals, refers to all of these people collectively. The lowest grade is that of an examiner, which is also known as a marker or, particularly in the literature, as a 'rater'.

The term 'script' refers to the answers of a candidate; it is used both to refer to all the answers that a candidate gives to an examination, collectively, and also to individual answers to individual questions, depending on how the material is provided to markers: in some cases examiners receive the whole script, and in others answers are separated. The pre-marked scripts which examiners receive at the standardisation meeting are formally termed 'standardisation scripts' but are more frequently referred to as 'sample scripts.' 'Sample scripts' also refers to the sample of marking which an examiner provides to his or her supervisor. These scripts do not feature in these data, so 'sample scripts' always refers to the standardisation scripts. Additionally the term 'anchor scripts', which is used in the literature, may also be used.

The topic of inquiry

The process assumes the methodical application of the mark scheme to a candidate's script in a uniform way that is unvarying and is at odds with the constraints imposed by time and cognitive load. Yet examiners make acceptably accurate and consistent decisions, as monitored by the systems of the exam board. The question therefore arises, as to what is actually happening when examiners make their decisions. The investigation was guided by the following research question:

How do examiners make their decisions when assigning marks to essay scripts?

Examiners' decisions do not take place in a vacuum. Even if it seemed doubtful that they were occurring the way that the regulations and system suggest they do, the training and tools which are provided must play some part. If not, then this in itself is of interest, since providing them is a costly and complex matter for the Awarding Bodies. Therefore the overarching research question was subdivided into the following sub-questions.

- What decision-making behaviours do examiners exhibit?
- What training strategies are demonstrated by senior examiners and how do they relate to the decision-making behaviours exhibited?
- What foci of examiners' attention during reading can be identified?
- What tools do they use to make their decision and what is the role of the mark scheme?

A first phase of data collection used discourse analysis to address recordings of two kinds: a pre-standardising meeting and four standardising meetings (two live, and two online), two units for each of English and History were recorded and transcribed; in addition three examiners from different units supplied 'think aloud' data from their live marking, a total of between one and three hours each spread across the marking period. A second phase of data collection took place the following year, using a survey to gather data from a larger sample of examiners which was used to contextualise the findings of the first phase, and to provide a measure of validity for the conclusions drawn from it.

Structure of the thesis

The first two chapters set the problem in the context of the literature. Chapter 1 will outline some theoretical background material drawn from a number of fields, notably the cognitive psychology of decision-making. This theoretical framework informed the data analysis. After this Chapter 2 will look at the empirical studies which have looked at examining and judgement in the specific context of educational assessment, forming a more traditional review of the literature.

Chapter 3 sets out the research design, the methods of data collection and data analysis, as well as discussing the ethical implications of researching assessment, particularly within the context of an Awarding Body.

The findings are presented in six chapters, organised thematically. The first two chapters focus particularly on the training meetings, before moving on to specific principles

of decision-making, beginning with the mark scheme, as it represents the perceived process of judgement, and then considering factors which emerged from the data and the literature.

Chapter 4, therefore, examines the training meeting, characterising the differences between the different meetings and considering the training strategies which are employed by the senior examiners, particularly the interpersonal and affective aspects of the meeting. Some consideration is given to the differences between online and face-to-face meetings. Chapter 5 extends this discussion of the training meeting by scrutinising the model analysis of answers by senior examiners, and suggesting a typology for the comments which they make. This is then linked to the analysis of scripts which is evident in the 'think aloud' data from individual examiners.

Chapter 6 considers the role of the mark scheme in the process, examining the ways in which examiners interact with it, and how they interpret it. Chapter 7 considers the use of comparison in the decision-making process, establishing it as a major tool in examiners' thinking, and laying out the possible comparative material. In particular it assesses the use of the anchor or sample scripts, what their characteristics are, and how they are used in the examining process.

In Chapter 8 the role of heuristics is discussed. A number of unofficial heuristics are offered by the senior examiners during training, which are then used by examiners during their marking. The relation of these, and other aspects of decision-making, to the more formal cognitive heuristics theorised by Kahneman *et al.* (1982; revisited in Gilovich *et al*., 2002) is then examined, and the evidence for the use of these heuristics is laid out.

Chapter 9 presents the data from the second phase of data collection, drawing together the themes considered in earlier chapters, and discusses the results of the survey in the light of the earlier findings.

Finally, Chapter 10 presents the conclusions of the study, relating the findings to the original research questions, and to the theoretical background which framed the data analysis. It considers the limitations of the study, and its contribution to the understanding of the examination process. It also suggests some potential implications of the findings.

Chapter 1: Theoretical Framework

The review of the literature has been structured in two sections. This chapter will examine the theoretical material that has informed the development of the research questions and the analysis of data, drawing on a number of different fields to construct a framework under which to consider the decision-making processes of examiners. The following chapter will review empirical studies of examining.

Initially I will consider the principles of cognitive load and limited working memory which frame the assumptions of the research and use them to suggest that examiners cannot meet the assumptions made by Awarding Bodies and the regulator. A potential explanation for the apparent ability of examiners to circumvent these restrictions might be supplied by theory of experts, who demonstrate extended working memory. The literature on expertise is therefore considered to establish its relevance or otherwise.

The principal cognitive theory of decision-making examined is that of heuristics and biases, as established by the work of Kahneman *et al* (1982; updated in Gilovich *et al*., 2002), which posits the existence of cognitive 'rules of thumb' which allow the brain to make short cuts during decisions; this can lead to bias, but does not intrinsically do so. Three of the heuristics (*representativeness*, *availability*, and *anchoring and adjustment*) stem from the original work; these are supplemented with the *affect* heuristic, posited by Slovic *et al.* (2002). This work was developed by Kahneman & Frederick (2002) to a cognitive model of decision-making that suggested a dual system (System 1 and System 2) where System 1 was the automatic heuristics-governed decision-making process and System 2 was rule-based, slower, rational decision-making. Kahneman & Frederick characterised System 2 as being one which supervises and monitors the intuitive judgements of System 1, with the assumption that 'an intuitive judgement is expressed overtly only if it is endorsed by System 2' (2002:57).

Some theory on decision-making in groups is also considered, given the context of the examination standardising meeting from which most of the data in this study is drawn, for potentially relevant material. Finally the model of connoisseurship is also considered as a potential explanation for the ability of examiners to make judgements which are subjective yet still considered 'accurate'.

Cognitive load and working memory

Models of cognition sometimes use a metaphor of the brain as a computer, 'processing' data into outcomes. The word 'attention' is commonly used to refer to 'selectivity of processing'

(Eysenck & Keane, 2000:119). While models of cognition differ about the amount of processing power to which the brain has access, there seems to be a consensus that limits are placed on cognition by the requirement of 'attention': either processing is seen as only occurring while attention is focused on the problem, so processing is very limited, or it is seen as a process operating on many parallel paths simultaneously, where the end stage involves channelling through a conscious focus of attention, leading to a bottle-neck situation. In either case there is a limit placed on cognitive processes, which is a key assumption of this study. If we can 'barely attend to more than one object at once' (Marois & Ivanoff, 2005: 296) then the objects to which we do attend become significant; the features to which examiners pay attention, as demonstrated by their being mentioned, will be a focus of data analysis in this study.

A variety of studies focussing on auditory and visual processing have established limitations in the number of stimuli and the number of tasks with which individuals can cope (Marois & Ivanoff, 2005). Spelke *et al.*, however, rejected these limitations, arguing on the basis of an experiment in which they trained two volunteers to carry out two different information processing tasks simultaneously, that 'people's ability to develop skills in specialised situations is so great that it may never be possible to define general limits to cognitive capacity' (1976:229). Examining would certainly fall into the category of a 'specialised situation', and the effect of training is of interest. Various studies since have sought to replicate, echo or challenge these findings. Broadbent (1982) argued that there were always some signs of cross-task interference when the data from such studies were investigated closely: so limits do exist, despite training. Eysenck and Keane (2000) suggest that while a complex task may initially require the application of a number of individual processing resources, after practice the number may be reduced. However, more recent neurological studies have suggested that although cognitive limitations are not absolute, the training required to enable, for example dual-task multi-tasking, must be prolonged and extensive (Dux *et al.*, 2009). Dux *et al*.'s study also sought to find the neurological mechanism whereby training could make for more efficient multi-tasking; they found evidence to support Poldrack's (2000) account that the 'flow of sensory motor information from each task would be progressively routed away from slow deliberative processing in prefrontal cortex, thereby bypassing the neural locus of multitasking limitations' (Dux *et al.*, 2009: 131). While neurophysiology is well outside the scope of this study, this is interesting as it appears to support the theory from the field of psychology of migration from System 1

to System 2 judgements, discussed below, in that it moves from deliberate, attention-rich cognition to a faster, more automatic cognitive process.

The limits on information processing are typically related to the limits of short-term memory (Lavoie & Grondin, 2004). Working memory, as conceived by Baddeley and Hitch (1974), consists of three components: the central executive ('attention'); the phonological loop (auditory component) and the visuo-spatial sketch pad (spatial and visual component). Each component is limited in its capacity and is independent of the others. The question of what is held in the working memory is key to an examiner making a judgement based on the combination of rubric and essay. To perform complex cognitive tasks, people must maintain access to large amounts of information. In terms of cognition, 'complexity' requires relatively little; it includes reading sentences in order, where the reader needs access to, for example, previously mentioned subjects and objects to resolve references to pronouns (Ericsson & Kintsch, 1995). The retention of both the contents of an essay, the contents of the mark scheme and the relevant features of a sample essay go far beyond this 'complex cognitive task' of reading, and must be performed in addition to it.

Ericsson and Kintsch (1995) sought a model of working memory that could reconcile subjects' 'limited working-memory capacity in laboratory tasks' with the 'greatly expanded working-memory capacity of experts and skilled performers' whom they characterised in the form of chess masters and others who perform 'complex cognitive tasks' which must surely include the awarding of marks to examination essays (1995:3). They proposed a 'Long Term Working Memory' which enabled an extended range of pattern recognition, with patterns kept in long term memory, but easily accessible by means of cues which sit in short term memory; a theory which sits well with the improved memory capacity of expert chess players which will be discussed in the next section. It also coincides with the techniques used by examiners of short answer questions (Suto & Greatorex, 2008b). They draw on the superior memory performance of, for example, waitresses in restaurants, who use table positions as cues for recall of complex drinks orders. They review a number of different areas of research, and conclude that in addition there is ample evidence for the presence of the relevant information in Long Term Memory after the performance of a skilled task. They conclude that

> part of being a skilled problem-solver is to possess exceptional memory skills. These memory skills are always highly specific to the domain and can be acquired only through extended practice. They cannot be based on temporary storage in ST-WM because many of

> the tasks discussed require much greater storage capacity than is available in STM (Ericsson & Kintsch, 1995:43).

This brings us to the topic of experts: individuals who possess domain-specific skills acquired through extended practice.

Experts and Expertise

The field of expertise and expert performance would seem to provide a potential precedent and explanation for individuals undertaking tasks which exceed the usual cognitive limits and stretch the performance of short-term memory. It is a field which has typically focused upon high performers in the fields of music, sport, medicine, and chess, though not exclusively. Researchers in this field have sought to explain the elite achievement of individuals who demonstrate superior cognitive processing, memory performance or performance of other complex tasks. This superiority is confined to specific aspects of the domain under consideration, rather than being generalisable to other experiences, as, for example, Djakow, Petrowski & Rudik's (1927) research with chess experts, whose memory performance was limited to chess positions. Most considerations of experts work on the analogy of elite performance in music and sport. The high levels of practice and the time taken to acquire expert performance are generally cited as the defining characteristics of an expert; ten years of consistent practice is considered a minimum in many fields (Ericsson, Krampe & Tesch-Römer, 1993).

This is *not* a criterion which examiners can fulfil. Many examiners do not have ten years of teaching practice, let alone of examining experience. Nor can the intensive periods of examination marking, for a few weeks twice yearly, be considered as the 'consistent' practice required; although they are of sufficient intensity they fall far short of the required duration. Between examination sessions examiners may mark student essays but this is a very different activity. On this basis one might reject the label of 'expert' for the decision-making processes of examiners. However, there are some aspects which are worthy of further examination, given that there are parallels.

The other common definition of an expert which is cited is that of an individual who displays 'special skill or knowledge derived from training or experience' (Merriam-Webster, 1979). This much more open definition would apply to examiners. Shanteau & Stewart adopt this definition and claim it to be consistent with their view that 'expert judgement applies in situations where there are grounds for saying that some judgements are better than others' (1992: 95). Are there grounds for accepting that this applies in the context of examiners'

decision-making? It is certainly the principle upon which the system in England, Wales and Northern Ireland is founded; the judgement made by the Principal Examiner is by definition a 'better' judgement than that of the ordinary examiner. All other judgements, therefore, can be ranked by their closeness to the definitive one, and indeed, all examiners are ranked according to their closeness to the accepted standard of judgement by means of their grade. Examining is a process requiring subjective and professional judgement; there is no absolute or correct mark that a script should receive, except that provided by the Principal Examiner's decision. A less technical definition is attributed to Niels Bohr, the physicist, who declared that 'an expert is a person who has made all the mistakes that can be made in a very narrow field.' In an educational context, learning from mistakes can be seen as the most profound educational experience, and examining, particularly of one module of one subject, is a very narrow field indeed.

K. Anders Ericsson, in his introduction to *The Cambridge Handbook of Expertise and Expert Performance* (Ericsson, Charness, Feltovich & Hoffman, 2006), sums up a number of ways of conceptualising expert performance. The two which seem most relevant are 'expertise as the extrapolation of everyday skill to extended experience' and 'expertise as reliably superior (expert) performance on representative tasks' (Ericsson, 2006:11–13). The term 'representative' is key: experts only function better on their specific area of expertise. For both conceptualisations, experts have no better performance on, for example, random memory tasks using the same materials as the representative tasks (Chase & Simon, 1973). Increased memory performance in the representative task is seen as dependent upon a large body of patterns which have been acquired through long years of practice, that is, it functions as recognition of an already known pattern. The 'reliably superior' judgment is an interesting criterion; two studies (Camerer & Johnson, 1991; Bolger & Wright, 1992) found that there are domains where experts perform no better than novices, and both the reliability and validity of their judgements can be flawed. Performance does not automatically increase with experience, but dedicated practice activities can and do improve performance, as Ericsson, Krampe & Tesch-Römer (1993) found in their synoptic review of the role of practice in expertise. It would be easy to characterise the training activities of the standardisation process as 'dedicated practice activities', and indeed they do fulfil the other criteria which Ericsson, Krampe & Tesch-Römer require:

> 'the design of the task should take into account the pre-existing knowledge of the learners so that the task can be correctly understood after a brief period of instruction. The subjects should

> receive immediate informative feedback and knowledge of results of their performance. The subjects should repeatedly perform the same or similar tasks' (1993: 367)

However, it is not difficult to see that these criteria relate strongly to most kinds of successful training activities, and it is unsurprising that the standardising process, designed for training purposes, should fulfil them. These are not the only criteria which deliberate practice aimed at gaining expertise must fulfil; Ericsson, Krampe & Tesch-Römer also suggest that it must be a limited amount undertaken daily over a long period of time – as much as 10,000 hours over a decade (1993:394), which is clearly not a characteristic of examiner training.

In many ways the system is designed to obviate the need for expert examiners: the provision of rubrics, model grades and a rule-based strategy for making decisions is intended to create a system which does not rely on expert judgement. It could be argued that the system requires just one true judgement to be made: that of the chief examiner who decides where the calibration of this year's standard sits. The intrusion of others who consider themselves expert – bearing in mind that the majority of research into expertise in judgement and decision-making shows that greater experience and expertise makes an individual more confident but not necessarily more accurate (Shanteau & Stewart, 1992) – may in fact be unwelcome. However, consider again Shanteau & Stewart's statement that 'expert judgement applies in situations where there are grounds for saying that some judgements are better than others' (1992:95). This reflects the situation in examining perfectly: those judgements which are better are those which are more aligned with those of the chief examiner.

It is clear that the individuals who mark scripts must in some ways be expert: they follow 'deliberate practice activities' (Ericsson & Lehman, 1996:273), they fulfil the criteria required by Ericsson & Kintsch (1995) for those who may display extended working memory by virtue of expertise, and they amply fulfil the dictionary definition. One of the key findings of psychological research into expert thinking is that some experts, such as chess masters, use pattern-based retrieval from memory to produce an appropriate solution to a problem (Ericsson & Lehman, 1996:275). This is supported by the findings of Suto & Greatorex (2008b), discussed in the next chapter. Pattern-matching seems to be inappropriate in the current context, however, since the greater complexity of essay questions, answers and marking, produces few if any limits on the numbers of ways answers may be made. However, it could be that the role of annotation in the marking of essays may enable a form

of pattern-recognition in terms of providing a schematic representation of the contents of the essay, expressed in terms of the rubric, although annotation is becoming less frequently used as marking migrates to an online context.

There is a limited but substantive body of work on the concept of expertise among historians, beginning with the work of Wineburg (1991), a relatively late beginning in comparison with research on expertise in areas of 'well-structured' problems such as chess, medicine or formal logic. Voss & Wiley, in their overview of the research, characterised an expert historian as someone who has 'a general and a specialized knowledge of history as well as facility in the skills of historical research and writing' (2006: 569); the domain of history is generally

> 'concerned with ill-structured problems that have a large amount of potentially related information, and different experts may approach the same issue differently, depending on the expert's theoretical background, related knowledge and other factors. Such solutions are usually verbal arguments, which typically do not have right or wrong answers, but the answers may vary in relative acceptability' (2006:570).

I quote at such length because this definition seems to apply in some ways to the 'problem' of decision-making during examinations also: there is a large amount of potentially related information, and whether it is an 'ill-structured' (more than one possible answer, having an agreed-upon solution) or a 'well-structured' (one answer, readily identified constraints, an agreed solution) problem is debatable. Although examiners are not seeking to arrive at a verbal argument, given that their decision is expressed in terms of a single numeric value, that value is attached to a description of attainment, and the process of coming to the decision often takes the form of a verbal argument and justification, whether as annotation or oral commentary, during training or later (as will be seen in later chapters).

Drawing on the body of research into history expertise, Voss and Wiley postulate ten 'Characteristics of History Experts' or 'CHEs', some of which seem relevant to the exercise of arriving at a judgement of an essay. Under the heading of 'Obtaining Information' they give (CHE1) 'historians evaluate sources' judging their usefulness and authenticity, (CHE2) 'experts use at least three heuristics in their analysis of sources, corroboration, sourcing and contextualization', (CHE3) 'when analyzing sources historians develop mental representations of the events and activities discussed in the text (situation models) and also generate subtext' (2006: 571–2). In the category of 'Reasoning and Problem Solving' they

include (CHE9) ‘evidence or justification for a claim or conclusion is usually verbal...[making] use of “weak” methods of reasoning and problem solving such as analogy, decomposition, and hypothesis or scenario generation’ and (CHE10) the attempt to deal with the general absence of control groups by the use of ‘counterfactual reasoning’ (2006:577–9). These characteristics may all aid the examiner in the ‘problem’ of reaching a judgement on an exam script, if one conceptualises that problem as ‘what mark would the Principal Examiner give this essay?’, a question which requires the examiner to draw on the sources of the script itself, the mark-scheme and the sample marked scripts which they have received. They generate mental representations of the essay, either by annotation or by commentary, as discussed in the review of the empirical literature and in Chapter 5 below. They create hypotheses concerning ‘what if the script was like this’ (as considered in Chapter 8) and they use as many methods of reasoning as they have at their disposal to solve their ‘problem’. Are examiners of history, who could be considered as expert historians, by virtue of their study of the subject (although Voss and Wiley suggest the first year of the PhD as the time when historians begin to become ‘expert’), demonstrating similar or the same characteristics in their examining? There is no literature that considers experts in English literature, but it is reasonable to assume that some, at least, of the same characteristics apply, particularly CHE1, CHE2 and CHE9, in dealing with the analysis of literature. It is possible, therefore, to see English and history teachers as expert at qualitative analysis of texts, which they can then apply to examination marking. This does not fit entirely with the strong evidence which suggests that expertise is so domain-specific as to be entirely untransferable, as discussed above, but it is worth considering.

If we consider expertise to be essentially equal to the ability to circumvent processing limitations, then it is possible that examiners may be ‘experts’ in what they do, and are therefore able to sidestep the constraints of working memory and attention which have been suggested as being incompatible with the examination marking system working as it has been designed to. However, the literature on experts and expertise demonstrates that expertise is a very domain-specific notion, and that though experts in different domains may have characteristics in common, they have many more characteristics which are not shared. There is no established method for experts to circumvent normal processing limitations, so the literature becomes relevant but not directly so, as one must still look for an explanation of how these particular experts function in the specific circumstances. In seeking for this explanation this thesis draws on a wide area of decision-making and psychological theory, which will be considered in the rest of this chapter.

Connoisseurship

The current study seeks to explore the processes by which examiners can continue to make accurate judgements through subjective means. There is a well-established body of literature on subjective judgements, or even expert subjective judgements, which uses the concept of the connoisseur. This literature is particularly prevalent in the context of art and music criticism, but has not been widely used in reference to examination judgements. Since the mid 2000s there has been an increased use of the concept in reference to teaching and teacher assessments, drawing on fine arts traditions, and emphasising the art and craft aspects of teaching and of subjects. Robbins (2008) suggested connoisseurship as an alternative model for school assessment, drawing on the tradition of connoisseurship as used by assessors in, for example, Associated Board of the Royal School of Music grade examinations. He suggests three characteristics of connoisseurs:

> the person is qualified to make judgements
> the exercise of critical faculties is based on knowledge and experience
> there is an ability to make comparisons in relation to perceived qualities. (Robbins, 2008:5)

The first two characteristics are both also demanded by Awarding Bodies in their recruitment of examiners, and the role of comparison will form a major theme of the data in this study (Chapter 8); the question of 'perceived qualities' will also arise (Chapter 9). Robbins also suggests that:

> 'Authentic connoisseurship is achieved through a process of induction into the community of assessors; it appears to be an iterative process that generates hermeneutical understandings that are then used to mediate the norm referenced assumptions used as the basis for judgements' (2008:5)

This resonates with the idea of 'guild knowledge' which will be explored in the next chapter, that draws on the idea that individuals are apprenticed into a community of subject/assessment knowledge, an underlying understanding which they then draw on when making judgments. The main difficulty with Robbins' definition is the fact that he asserts that connoisseurship relies on norm-referencing, a direct contradiction of the supposed operation of current assessment on criteria-referenced principles. The recurring theme of experience, however, links connoisseurship back to expertise, and it is hard not to see the

connoisseur as a facet of the expert, working in a specialised area of qualitative judgement – a response to an 'ill-structured' problem, perhaps. Certainly this model is in keeping with the belief of some that 'educational assessment must be understood as a social practice, an art as much as a science, a humanistic project', and that the judgement of human beings can deal with the 'things we cannot tell' rather than concentrating on the metrics of assessment (Broadfoot & Black, 2004), a sense which resonates with the consideration of holistic assessment in the next chapter.

Discourse psychology and pragmatics

In addition to these models of cognition and decision-making, it is necessary to consider discourse psychology. Graesser *et al*. stress that it is 'intertwined with virtually all cognitive functions and processes, including memory, perception, problem solving and reasoning" (1997:165). Examination scripts are, no less than any other written utterance, primarily a form of communication, albeit one which operates within a particular set of rules, with certain expectations appended to it. The ways in which readers comprehend writing is relevant, therefore, in that the examiner's comprehension of the examination script will follow normal reading processes, constrained by time pressure and the framework provided by the rubric.

Graesser *et al*. elucidated and developed the Gricean maxims of pragmatic communication, which focused on quality (i.e. truthfulness), quantity, relevance and manner, into seven principles which govern clarity of communication:

1) *Monitor common ground and knowledge*. The writer should keep track of words, ideas and entities that the reader already knows. If something new is being introduced, it should be signalled syntactically and embellished with adjectives, phrases or examples.

2) *Use discourse cues to distinguish "given" versus "new" information*. For example, the given information is typically included in the subject noun-phrase of a clause and the first clause of multi-clause sentences, whereas the new information is in the verb-phrase and additional clauses.

3) *Use discourse cues to signal important information*.

4) *Make true claims about the situation model under consideration*. In expository text, claims should be true about the world in general.

5) *The incoming sentence should be relevant to the previous discourse context.* New topics, subtopics, and episodes need to be flagged with discourse cues, such as subtitles and transitional phrases.

6) *The order of mentioning events should correspond to the chronological order of events in the situation model*.

7) *Statements should not contradict one another*. (excerpted from Graesser *et al*., 1997:172-3)

If one or more of these principles is flouted by a communication, then the recipient will judge its quality accordingly. Graesser *et al*. assert that they are 'automatized and unconscious in the minds of most readers, at least those who do not work in a communication profession' (1997: 173). Given that some of the participants in the study will be teachers of English, arguably a 'communication profession', they may well be aware of the impact of pragmatics on communication, though that may or may not be expressed consciously in a way that can be picked up by a researcher. In any case responses to scripts based on pragmatic reflexes may constitute part of the examiner's decision-making process. If violation of pragmatic principles causes difficulty for the examiner in terms of text comprehension, then it may result in their decision being harder to make, or in a penalty for the candidate.

To some extent pragmatic principles are codified within the rubrics supplied by awarding bodies, so that, for example, 'indicative content' might suggest information that was 'true' or 'relevant' in relation to the question. The use of criteria such as 'organise relevant material clearly and coherently using specialist vocabulary where appropriate' (Edexcel, 2005:11) or 'factually accurate', 'generally coherent in expression and cogent in development' (AQA, 2009:6), speak to the relevance of pragmatics in reaching examination judgements. The concept of 'Quality of Written Communication', which remains nebulous for many examiners, may provide a formal home for examiners' awareness of pragmatic principles, and forms a recurring theme during data analysis.

Heuristics and Biases and System 1 and 2

Kahneman and Tversky's theory of heuristics and biases provides an explanation of the ability of the human to make intuitive judgements when the cognitive load is beyond their rational capability, using three heuristics which involve assessing the information that is available, judging its representativeness against samples from their prior experience and then adjusting to fit the scenario currently in question. The 'biases' are where error creeps into, because of lack of availability, skewed recall of examples *et cetera*. This theory was developed during the 1960s and 1970s in a series of papers that sought to explain apparently

irrational decisions made by participants in psychology experiments in terms of intuitive judgements, based on estimation procedures. Heuristics and biases have been seen to fit with the concept of the 'cognitive miser' (Fiske & Taylor, 1984), that is that the human brain naturally develops systems which minimise the cognitive effort required in order to complete tasks or make judgements.

This field has become a large one, with evidence for many different heuristics produced by a variety of researchers. Those which may be relevant to the examining process are outlined below. The examples tend to be those where the heuristics lead to corresponding biases, since these illustrate the case more clearly, and it is in these situations where it is easier to identify heuristics at work. Biases often occur when cues for judgement are wrongly weighted in the judgement process (Kahnemann & Frederick, 2002). However, heuristics are a valid and well-used, automatic strategy in human judgement processes.

Availability The ease with which an example can be brought to mind alters the prediction of the frequency of an event, or of the proportion of the population. The statistical probability of an event is found to be less credible than the anecdotal, experienced example (theorised by Tversky & Kahneman, 1974).

e.g. An examiner marks an essay which should receive full marks. However, she might predict that very few or no essays should do so, having never seen an example, and give it fewer marks than it deserves.

Representativeness Items which appear similar are assumed to have the same characteristics, and an item which appears to fit into a group is assumed to have the characteristics of that group. This results in a situation in which 'some probability judgements (the likelihood that *X* is a *Y*) are mediated by assessments of resemblance (the degree to which *X* "looks like" a *Y*)' (Kahneman & Frederick, 2002: 49–50). Bias may creep in where, for example, the statistics of a sample are used to predict the characteristics of the population, even if the sample is too small to be 'representative' (theorised by Tversky & Kahneman, 1974). This is analogous to the concept of illusory covariance or the 'halo effect' (Thorndike, 1920).

e.g. A script whose content is worthy of an A grade may be under-judged because the handwriting and spelling are considered to be sub-standard. Equally a well-presented script may appear representative of a higher grade than its contents require.

Anchoring and adjustment This heuristic refers to the tendency to form initial judgements around a base point or 'anchor' and then adjust according to the situation and developing information, to reach a final decision. Using an inappropriate anchor and/or failing to adjust sufficiently leads to bias (theorised by Tversky & Kahneman, 1974).

e.g. The process by which an examiner sites an answer within a band (anchor) and then fine-tunes the mark up or down (adjustment) for the final judgement. Anchoring effects can easily lead to bias in examination: using the first of two essays in a script as an anchor for the second, for example, or using the initial paragraph or page as an anchor.

Affect This heuristic, more recently theorised, seeks to provide a role for the affective, as well as the cognitive, in decision-making. 'As used here, *affect* means the specific quality of "goodness" or "badness" (1) experienced as a feeling state (with or without consciousness) and (2) demarcating a positive or negative quality of a stimulus' (Slovic *et al*., 2002). Zajonc, one of the scholars whose work was a precursor of this theory, suggested that human decisions were often based on affect, before they then seek 'to justify these choices by various reasons' (1980:155). This heuristic suggests that an emotional response, perhaps unconscious, by the examiner to the quality of an examination script, could dominate the judgement which is made, which is then justified by reasoned argument (theorised by Finucane *et al*, 2000).

e.g. An examiner likes or dislikes a candidate's phrasing, handwriting, use of quotation, or virtually any other aspect of their writing, and this affective response dominates the judgement which is made. This is more damaging if the affective response is not consciously recognised, so that compensation may be made accordingly.

Kahneman & Frederick (2002) accepted the affective heuristic, and stated that it should replace anchoring as one of the three main heuristics by which intuitive judgements took place, since affect, representativeness and availability all rely on 'substitution' of one attribute for another, more difficult attribute which is the rational cue for judgement. Despite this, this study considered all four heuristics.

This acceptance was part of subsequent development by Kahneman & Frederick (2002) which produced a theory of a dual process system, where heuristics and biases evolved into System 1, while System 2 represented rule-governed rational judgements. Operations originally dealt with under System 2 may migrate to System 1 as an individual acquires skill and experience with situations requiring those operations. System 1

judgements made during the examining process may well use information that is not covered in the supplied rubrics, or they may use proxy information instead, in order to make judgements. Previous studies have looked at the strategies which examiners use, but they have not specifically looked at the match or lack of it between those strategies and the supposed rule-governed systematic judgement intimated by the existence and promotion of the mark scheme. Kahneman & Frederick characterised System 2 as being one which supervises and monitors the intuitive judgements of System 1, with the assumption that 'an intuitive judgement is expressed overtly only if it is endorsed by System 2' (2002:57). They present the Stroop test as evidence that the two-system structure is in place and working; participants are asked to report the colour in which words are printed. If the word is the name of another colour (e.g. 'red' printed in yellow ink) then participants hesitate and experience conflict, but make few errors, suggesting generally successful monitoring. The extent to which results from this simple experimental task are replicated in real world decision-making is questionable.

Others have rejected the division of judgements into these categories: Laming (2004) prefers to consider apparently 'System 2' judgements to be the cumulative effect of many much smaller intuitive judgements.

Decision-making in groups

Although a great deal of literature on individual decision-making exists, it is also potentially relevant to consider the more limited literature on decision-making in groups. During the standardisation meeting the examiners convene in small groups for training and to grade sample papers for practice. It is a public forum, in which voiced opinions are often tailored and influenced by those of others in the group. Christie and Forrest's book considered the grading meeting which takes place at the end of the examination process; they outlined a meeting of 'sweet reasonableness' (1981:35) in which statistical information and judgement coincided to make the grading decisions easy. They also posited a scenario in which the statistical and the subjective did not coincide, which they described as a 'contest' model of decision-making; in this situation they considered that it was entirely possible that the strongest personality would prevail. Alternatively 'a small compromise might be made. Yet successive small compromises could lead to a marked shift in standards' (1981:35): a problem which remains in the current system. The influence of personality on equal-status groups is a recurrent feature of the research into group decision-making.

Here I have considered the research which concerns itself with performance groups, as opposed to teams; the social psychological research considers teams to be longer-term groups (Kerr & Tindale, 2004), and despite the name given to them by Awarding Bodies, examination teams are ad-hoc groups who are usually newly formed and the members of which do not have any prior knowledge of each other.

The findings of Bettenhausen & Murnighan (1985) suggest that in groups which have been newly formed there is a tendency for the group to establish a new norm; the interaction between members causes them to either 'tacitly revise their beliefs about appropriate action, implicitly agreeing with the direction being taken by the group, or overtly attempt to pull the group toward their own interpretation though challenges to the implied norm' (1985: 350). Their study principally used observation of problem-solving among groups given tasks of the team-work building kind, and draws for the most part on behavioural norms, some of which are relevant to the formation of a working team from a group of examiners who do not know each other, but is less relevant to the cognitive processes of decision-making, other than to reinforce the suggestion above, that individuals within groups conform to what they perceive as the general opinion. They also recognise that most groups have to deal with some kind of challenge to their emerging norms: 'a threat or challenge arises when the group's recent agreement or current proposals fall outside what at least one member considers to be appropriate and that person is sufficiently motivated to act' (Bettenhausen & Murnighan, 1985: 356). This is the place where Christie & Forrest's consensus and contest models intersect; in regard to its applicability to the examiners' training groups, the question arises as to whether members are 'sufficiently motivated to act'. Examiners' teams are not egalitarian in nature; they all have a more senior member, or leader, who governs and guides the process, which may prevent these negotiations occurring. Although early research into group decision-making tended to focus on preferences (Kerr & Tindale, 2004), a shift towards consideration of information sharing has influenced research for the last two decades, following the somewhat counterintuitive finding of Stasser & Titus (1985) that groups did not share information optimally, and tended to only use information which was widely shared among the group, ignoring that which was not, a finding which has been reconfirmed many times since in the literature (Kerr & Tindale, 2004). Groups will focus on and discuss the information which they have in common, instead of attempting to uncover new unshared information which might be relevant. This is potentially relevant to discussion by examiners' teams in standardising meetings, in which the massive amount of information which might be relevant, and is available in theory, is

unlikely to be accessible to all members at any one time: each member of the team will recall certain items of relevance.

Similarly to the ways in which groups share information, Davis (1996; also Davis *et al.*,1993) found that groups gave more weight to the opinion of members when that opinion was closest to that of other members; the more discrepancy, the less weight attached. This, like the creation of norms and the sharing of information, suggests that groups will be less-than-optimal decision-makers, ignoring outliers. It would seem intuitive to think that group judgements of examination scripts should be more powerful, more 'accurate', but the theoretical literature suggests otherwise. It might also be considered that group decision-making would protect against the bias of individuals, but given the principles suggested above, this seems less likely, and indeed various researchers have found that the relationship of group decisions to individual bias is a complex matter, and can both attenuate and exacerbate its influence (Kerr *et al.*, 1996; 1999).

It is worth noting that some research has found that groups using electronic communication work at a disadvantage, particularly if close communication is required for the task, or if the members are unfamiliar with the electronic technology (Hollingshead *et al.*, 1993; Straus & McGrath 1994); the online standardisation meetings which are present in the data certainly demonstrate the latter of these two problems.

The extent to which this group decision-making literature is relevant to the data is debatable; to a large extent most of the data does not represent a group coming to a single decision during the standardising meeting, but an individual coming to a decision of their own in a group context. Consensus is not required, except post-hoc, after the 'correct' judgement has been revealed. As suggested above, the groups in this study are not egalitarian, instead being run as part of the hierarchy of the Awarding Body. This not only gives greater weight to the information and preferences expressed by the senior member of the team, but also means that individuals are operating in a scenario where their individual judgements are under scrutiny, in a high-pressure environment of the sort which has been found to prejudice performance. However, there is one group in the data who do tend towards the group decision and consensus model, and it is worth considering all meetings in relation to this area of literature.

Conclusions

Examination judgement, particularly post-meeting, is a matter of individual cognition, as has been the majority of the theory considered here. The findings chapters will discuss the data in the light of these ideas taken from psychological research into judgement and decision-

making. The final conclusions of this thesis will seek to establish which of the theoretical approaches are useful in the field, and attempt to create a synthesis. It is important to acknowledge, however, that one of the principles of this study is that individuals make individual decisions, and that what is true of one individual may be true of another to a greater or lesser extent. The ideas outlined here are usually not mutually exclusive, and they provide a useful theoretical framework within which to consider the data. Theory requires a context however, and the next chapter will consider the educational assessment context and empirical studies which have looked at examining.

Chapter 2: Examining the Literature

The prevailing tone of the engagement between English teaching professionals and assessment can be easily captured by the title of Brian Jackson's book *English versus Examinations* (1965); there has been a considerable amount of research into the assessment of English in England, Wales and Northern Ireland, but much of it has been devoted to refuting the validity of the assessment regimes imposed by the government at various times. This antagonistic attitude has coloured much of the resulting literature, but it has also been a stimulus for writings on the topic, even if they are not always based on empirical research. There is not a similar body of literature for the assessment of history, an impression confirmed by a hand-search of *Teaching History*, the professional journal, for the last ten years; any articles on the topic of assessment, and even an entire special issue, were focused on formative assessment rather than national examinations. However, the lack of subject balance has not troubled this literature review, as the subject-specific literature on English and assessment is hardly more than polemic for much of the time. Having dealt with relevant areas of theory above, this chapter focuses on existing empirical studies of examiners, which for the most part consider the decision-making processes of examiners without doing more than note their subject domains. The subjects under assessment will be considered when it appears relevant in this review, just as distinctions between history and English examining will be made where they are relevant later in the thesis.

Searches were carried out for combinations of the key words 'examin*', 'think*', 'cognit*' 'mark*' and 'judg*', in the British Education Index, ERIC and OpenSIGLE. A hand-search was made of *Assessment in Education: Principles, Policy and Practice* for the last ten years, all issues of Cambridge Assessment's *Research Matters*, and of the Cambridge Assessment list of publications. Inclusion and exclusion of the search results was judged on title and abstract. In addition citation searches were carried out for, as well as following up references from, papers found by the search. Much of the literature on examiners' thinking comes from the field of teaching English as a Foreign Language (EFL); where relevant this literature has been incorporated into this review.

Research on the assessment of essays in particular is more widespread in the EFL context in which holistic rating of essays is a common assessment technique. Hamp-Lyons (1991) demonstrates some of the reasons that literature from EFL contexts does not always assimilate easily into discussion of A level examinations; such as the fact that for her 'readers' (as she calls markers) there is no guidance for content, or that they may be 'unable to judge other aspects of the essay because they are bogged down in technical language'

(137) on a specialist topic such as engineering with which they themselves are not familiar. This renders some aspects of the research irrelevant since A level examiners are concerned with content as well as communication, and are to some extent at least, expert in that content, although, provided that the different context is borne in mind, this can be a fruitful source of literature on examiners and examining.

However, the main focus of this chapter is on studies of external examiners of secondary schools in the UK, in order to maximise their relevance to the current study. These studies are relatively sparse, although the research conducted by Cambridge Assessment is a notable exception, and has provided much of the material for this review. There are undoubtedly other studies which have been carried out internally by exam boards, but never published in the academic literature, to protect commercial data or because it would be impossible for their research context to remain anonymous upon such publication. Some of these studies have been published as reports by exam boards, or by OfQual or its predecessors, such as the Secondary Examinations Council's publication of Daugherty (1988), without anonymising the data; where I have been able to obtain such reports, which are rarely cited and more rarely available, I have reviewed their contents without including identifying aspects of the research contexts. Some of the research deals with the assignment of grade boundaries after the numeric scoring of essays; although the judgement processes involved are distinct, they are analogous activities, and such studies have been included, although the context has been made clear.

The following review is divided into five sections, which consider the use of essay assessment, three aspects of examiner thinking (cognitive processes, factors in decision-making and internal frameworks) and the limited research on the training meeting.

The trouble with essays

Thompson and Bailes (1926) provide one of the earliest studies, if not the first, to consider the difficulties associated with the reliable marking of the essay form. Their simple experiment required seven different markers to award marks on a scale of one to five to 50 short essays. The markers were in complete agreement on just one essay. Another essay was awarded each possible mark by at least one grader; a further nine scripts received a spread of marks from the raters which covered four of the possible five outcomes. The authors conclude that 'the correlations obtained by judges of essays are fairly satisfactory, *considering the great difficulty of the task*...the present correlations by their *comparative* [to two markings of the same intelligence test] lowness emphasize the fluctuating and

subjective nature of judgement on an essay' (Thompson & Bailes, 1926: 91; italics in the original). It is natural that much of the research on examining has followed this article in being concerned about this difficulty, as measured by varying inter-rater reliability. Many later articles take as read that essay-based measures are problematic forms of assessment: Coffman, for example, cites in his opening sentence the 'well-recognised limitations of essay tests with regard to reliability' (1966: 151) and suggests that they should be reserved for 'reference' tests, with the supposedly more objective multiple choice carrying the weight of assessment. It is true that he was writing in the US context, which is very different from the UK, in its reliance on standardised multiple-choice and rejection of constructed response assessments – the reverse of the UK school assessment system. However, despite their widespread use and acceptability in public examination, the limitations of the reliability of essay-based examinations is also taken for granted in many British studies, so that between the publication of Hartog and Rhodes's large scale study in *An Examination of Examinations* (1935) and Cox's review of the literature on assessment in higher education in 1967, there was no attempt to challenge the accepted view that 'examinations calling for the marking of long written essays are extremely unreliable' (Cox, 1967: 292), nor has there been one since. It is perhaps inevitable that research into examining has, therefore, focused on the reliability of marking and the reduction of inter- or intra-marker variation, even up to the present day. No attempt has been made to restore the reputation of the essay as an assessment tool, yet if marking them were as unreliable in the practice of public examination as it is described in the literature, then both the regulator and the Awarding Bodies would be in uproar. It is perhaps attributable to this acceptance that essay assessment can be very unreliable that Awarding Bodies have careful systems in place to monitor the decisions of examiners and to ensure that they operate with acceptable levels of reliability.

Yet it is also widely, if usually tacitly, acknowledged that there are some subjects, or parts of subjects, for which 'objective testing' (that is, the counting of simple correct answers to provide a total score) is unsuitable, and for which the measure has to be 'direct qualitative human judgment' of the quality of student work (Sadler, 1987:192). In a later study Sadler enumerated the characteristics of qualitative judgement, including the statement that there is often no 'independent method' to confirm whether or not a judgement is correct at the time of the decision, and indeed that 'it may be meaningless to speak of correctness at all. The final court of appeal is to another qualitative judgment' (1989:125). This is yet another difficulty with essays: whether a mark is 'correct' or otherwise is another qualitative judgement. In many ways the problem facing the examiner is not 'what is the mark which

this essay deserves?' but the related yet different question: 'what mark would the chief examiner give this essay?' To the former question there may be many different answers as judged by different readers, but to the latter there can only be one correct answer. It is upon this foundation that the English and Welsh essay examinations stand, and it requires absolute trust in the judgement of the Principal Examiner. In other examination systems, particularly in Higher Education, the 'substantial degree of subjectivity' in essay marking, particularly in the arts and humanities, is offset by using double-marking (Partington, 1994), which is not true for the vast majority of scripts in A level examinations.

Given that the limitations of assessments based around essays have long been recognised there are surprisingly few recent reviews of the examination of such assessments. One of the very few studies to deal with essay based answers is Sanderson's (2001) unpublished PhD thesis dealing with A level examinations in Sociology and Law. A primary aim of his study was to 'challenge those elements of the process which are accepted as "given"' (2001:18) including the contextual relationships, cultural norms and personal backgrounds of examiners. Sanderson characterised examining as a specialised form of problem solving, and sought to understand the marking process by considering all its stages, an approach which this study seeks to emulate. His major contribution was the creation of a theoretical framework which situated the use of heuristics, as described above, within a model of 'communities of practice' (Wenger, 1998), a theoretical approach then adopted by Crisp (2008a), and which has informed the theoretical framework of this study. The extent to which the communities of practice model applies to teams of examiners is debatable, however, given that their 'community' gathers for a single day, and has limited interaction even within that day. Sanderson's research will be further considered at the appropriate points below.

Having acknowledged the challenge of assessing essays it is essential, however, to also acknowledge that in the two subjects under consideration in this thesis, essay writing is a central skill. Harris argued that for history, essay writing is an integral part of the subject: 'without presenting an opinion and supporting it with substantiated argument, our subject scarcely exists' (2001: 13). The overall sense conveyed by Marshall in the opening chapters of *Testing English* (2011) is similar: continuous writing, of extended argument, is the most 'authentic' form of English assessment. To say that examination by essay is problematic is not to say that it is not still the best form of assessment in the subjects in question.

Holistic and analytical marking

Another of the presumptions behind the system of criteria specified on mark schemes, and the division of mark allocation into different Assessment Objectives, is that by doing so, examiners' decisions can be made more reliable, and the basis on which the assessment decision is made is therefore more transparent. Analytical assessment with carefully specified criteria has become the norm in the UK since the beginnings of the trend in the 1980s (Marshall, 2011), when Awarding Bodies started the movement away from more traditional norm-referenced examination. The motivation behind such a move was summarised by Peter Kimber of the Scottish Examination Board in a talk in 1984 in which he joked:

> we in the big wicked exam board are often accused of operating a secret society whose judgements are absolute and almost unquestionable but whose criteria are not explicit (Kimber, 1984, cited in Daugherty, 1988).

It is hard to conceive of a scenario in which Awarding Bodies could return to holistic marking for the examination of A level essay subjects. This practice of using detailed criteria is not without controversy in the literature, however, and Sadler has written extensively on the topic, although more frequently in the context of higher education than that of schools; he has challenged the idea that 'criteria-referenced' assessment is any more clear or any less subjective than holistic marking (Sadler, 2005). Sadler has previously suggested that true criterion-referenced assessment relies on objective testing of the presence or absence of said criteria, a thing distinct from the qualitative judgement of essays (1987). For Sadler qualitative judgement is 'not reducible to a formula which can be applied by a non-expert' (1989:124); while the creation of analytic mark schemes may not intend their application by such a non-expert, it certainly presumes the reducibility of the qualitative judgement.

Sadler considers extensively the characteristics of qualitative judgement. His arguments will be considered here in some detail as they are deeply pertinent to the examining process as considered in this study, and to the very question of whether analytic marking of some types of assessment is possible. He makes five main points, which are briefly summarised as:

1. Multiple, interlocking criteria are used which 'amounts to more than the sum of its parts. Decomposing a configuration tends to reduce the validity of the appraisal.'

2. At least some criteria are '*fuzzy* rather than *sharp*'; that is there is a continuous gradation from one state to another, rather than a discontinuity between two states.
3. A relatively small subset from a large pool of potential legitimate criteria are used at any one time.
4. There is often no independent method of confirming the 'correctness' of the judgement. (considered above)
5. 'If numbers (or marks, or scores) are used, they are assigned after the judgment has been made, not the reverse.' (Sadler, 1989: 124–5)

It is the first point which is most obviously relevant to the question of holistic versus analytical marking; the interaction of the multiple criteria which go into making a qualitative judgement make it problematic to separate them and weight them appropriately. Hamp-Lyons, in an overview of assessment research in the EFL context, supports this, suggesting that there are 'inherent qualities of written text which are greater than the sum of the text's countable elements' (1991:79). Indeed, Sadler mentions under the third point that even expert judgement makers are unable to specify all the relevant criteria which they may be taking into consideration, and not all criteria will be of use in any one case; he coins the term *metacriteria* to describe the criteria used to decide which criteria are of relevance. Sadler conceives of qualitative judgement as a complex and expert procedure; the analytical mark schemes demonstrate that this complexity is not lost in the move from holistic marking. The challenge that the sum of the parts of a qualitative judgement does not add up to the whole is one which resonates with the literature on expertise, and which will later be seen in the data of this study, but is hard to substantiate.

Qualitative judgement as described by Sadler cannot take place in a vacuum. There is a substantial body of literature which stretches back to the 19th century (Edgeworth, 1890) that documents the fallibility of holistic judgements, particularly those which are made by teachers; it is this body of literature which is frequently used to defend the UK external examination system against suggestions of greater use of teacher assessment, which might be considered to be more valid, relying as it does on constant accumulated knowledge of students and their attainment as opposed to a small number of assessment tasks. Sadler argues, however, that most of these studies measure similarity of judgement without

providing any reference material for calibrating that judgement – the 'standards' which he proposes.

The axiom that holistic judgement is unreliable is widely taken for granted. Vaughan (1991), working with EFL raters, was drawn to examine cognitive processes because the system she was working with relied on holistic marking, in which the examiner becomes to some extent the instrument, and therefore understanding that instrument is vital. The widespread use of holistic marking in English composition has been a stimulus for research in that context; such research is more recent than that based on examinations for A level or GCSE, where the widespread use of analytic mark schemes over the last thirty years seems to suggest the decision has been made. There is a crucial difference, however, in that the majority of certificated EFL assessments use multiple markers (Hamp-Lyons, 1990), as opposed to the single marker, with random sampling, of the A level system. Most recently, however, they too are moving to a more analytic approach, partly to support a move to computer-aided assessment (Way & de Jong, 2010).

The arguments concerning analytical versus holistic marking can often be reduced to the reliability and validity debate; analytical marking produces far higher reliability scores (Yao *et al.*, 2008), but it can be argued that the reduction of grading into analytical components does not adequately reflect the full complexity of qualitative judgements, and therefore skews the validity of the assessment away from the original intent (Sadler, 2009). In addition the work of Britton (1964) suggested that multiple 'impressionistic' markings of essays by a small team of examiners was not only faster than single examiner analytical marking, but was also markedly more reliable, even using the contemporary equivalent of a Principal Examiner as the standard against which to measure 'correctness' of grading. In addition to his team of three impressionistic markers, a fourth gave a mark for technical accuracy. Britton remained convinced of the strength of his experiment, even writing of it twenty-five years later (Britton & Martin, 1989).

The debate is of relevance to this study because it is one which is held in the examiners' minds, and which is discussed in one part of the data. The assumption has been that the division of mark schemes into numeric 'objective' analytical components has made for better assessment, in that there appears to be less reliance on the tacit knowledge of examiners (Sadler, 1987), and greater clarity in the citing of the criteria on which judgements are made. This assumption relies on the belief that examiners can make logical, rule-based decisions, using comparison of scripts to the mark schemes, something which is easily believable of subjects where the 'matching' strategies described by Suto & Greatorex

(2008b) are possible, but less so in extended essay subjects. It is possible, in fact, that by requiring two or three judgements on two or three Assessment Objectives to be made, all that is happening is that the cognitive difficulty of the task is amplified by two or three; natural qualitative judgements are restricted by the analytical requirements, particularly if examiners are unaware of the criteria under which they make their qualitative judgements (which they may well be, given the potential complexity and variance, which is considered in the next section), and are therefore unable to separate them. The collective decision to work on analytical rather than holistic grounds has been founded on the perception of improved reliability; an understanding of how examiners make their decisions can only improve the basis upon which the debate occurs.

Factors in examiners' decision-making

In the EFL context just as in A levels, the focus has been on the reliability. Vaughan rightly stresses the equal importance of evaluating the 'process' as well as the 'product' (1991:111) She used the 'think aloud' method to illuminate the decision-making of nine experienced EFL raters when marking the same six essays. She creates the following categories based on her participants' reading/marking strategies: the single-focus approach; the 'first impression dominates' approach; the 'two category' strategy (for example, grammar and organisation); the 'laughing' rater (whose articulations were dominated by her affect response, expressed as laughter); and the grammar-orientated rater (118–120). While her categories are interesting and demonstrate many of the characteristics of markers' thoughts that can also be seen in the current study, they are not mutually exclusive (the grammar-orientated approach would sit better as a sub-category of the single-focus approach, for example), and the extracts from participant data she quotes show characteristics of more than one style. One rater is not categorised, but comments by that participant elsewhere demonstrate a strong emotive reaction, which pairs well, if conversely, with the laughing rater. However, I agree with Vaughan that her data demonstrate that examiners react in different ways to essays and focus on different elements. Similarly Milanovic and Saville's study looked at the examination of holistically marked writing compositions in EFL, with the use of retrospective Verbal Protocol Analysis. They found an 'astonishing diversity' of factors being considered by the markers (1996:99), which emphasises once more the subjective nature of the process.

There is some evidence to suggest that certain aspects of scripts become key features in decision-making, such as 'application' and 'knowledge' in the grade boundary setting exercise for Geography A level undertaken by Crisp (2010a). The specific features

seem likely to differ from subject to subject, and may be directed, on the evidence from that study, by the key words of the particular Assessment Objectives, or comments from Senior Examiners. A wider range of features which played a smaller role was established through interviews with participants, corroborated by the verbal protocol analyses, including understanding of geographical concepts, geographical knowledge, the application of knowledge, 'sparks' of insight, consistency in performance, and the ability to write well, with the appropriate terminology, as well as targeting answers appropriately to questions. While some of these are very subject specific, it is clear that others have a wider application, and leaving aside the single subject context and the small number of participants, which Crisp acknowledges, these provide some indication of the number of features which examiners may take into account. It is also interesting to see that in comparing the behaviour of her participants in the grading as opposed to their individual marking, Crisp found that in grading the merits of a script became relatively more important than its weaknesses, in that more reference was made to them; this reflects an emphasis on positive rewarding of merits as opposed to 'negative marking'.

An interesting contrast to this is provided by language. While the language in which an essay is couched (or its 'Quality of Written Communication,' as it is termed in the literature of the Awarding Bodies and OfQual) is usually considered to be an entirely marginal matter, there is some suggestion that it is more than marginal in examiners' minds (as in Baird & Scharaschkin, 2002). Crisp (2010b) suggests that language becomes a more explicit or manifest criterion when there is some difficulty with the way in which a candidate has expressed themselves, which leads to problems with the interpretation of the answer by the examiner. She also suggests that when the length of a response does not correspond to an examiner's expectations, then that also becomes a factor in their decision-making.

Although it is a characteristic of a script which is most likely to be a factor in examiners' decision-making, there are some studies which have examined whether the characteristics of examiners play a role in their decisions. It has been suggested, for example, that less experienced markers are more severe than more experienced ones (Weigle, 1998; Greatorex & Bell, 2008). The gender of examinees has attracted some attention over the years, particularly in the light of gender gaps in examination attainment; Baird (1996) investigated whether the sex of the candidate caused marking bias in Chemistry and English Literature A level, and found no bias, as did the Scottish Examining Board (1992) in English and History, who used typed scripts to obscure the perceived ability to identify gender from handwriting. In contrast, Greatorex & Bell (2004) investigated the gender of examiners as

opposed to examinees, in English, Food Technologies and History GCSEs, considering both the biological sex and a measurement of gender as a socially constructed attribute. The subjects were chosen for exemplifying more male than female, equal numbers and more female than male examiners; they found no significant effects of examiner gender in any subject, although they did confirm the tendency seen in the EFL literature of more senior examiners to be more generous in their marking, in the English examination only.

Social and affective factors

Little attention has been paid to the role, or even the existence, of the affective in the marking of examination scripts. This may at least partly be because examination marking is meant to be wholly rational. It is one of the potential heuristics considered in the previous chapter, however, and some studies have noted both affective reactions and social perceptions of candidates by examiners. As usual, it has been more researched in the EFL context. It is considered under 'factors' not because it *does* affect examiners' decisions, but because it *may*.

Barritt, Stock and Clark (1986) found that less consistent essays were harder to mark and were therefore more likely to result in an awareness of the student author, entering into an imagined dialogue with the script in an attempt to understand what meaning was intended. This pseudo-dialogic interplay between examiner and examinee was also found by Delaney (2005) who reported an interaction with the writing that attempted to reach student thinking and which was 'almost entering into a dialogue with the student via the writing' (2005:5). It is possible that such interactions, seeking to understand the student author, are attributable to the ESL context, where poor language can impede communication. Vaughan (1991), also working in this area, investigated the holistic rating of English writing, as discussed above, and labelled possible emotional reactions to scripts as the 'laughing rater', one of five different reading styles she identified among markers, in which the affect reaction was the dominant characteristic.

Studies in the context of school examinations have been much more limited. Crisp looked at whether examiners showed 'social, emotional and personal reactions' during marking and reported a number of different reactions, including 'like, dislike, amusement [and] frustration' (2008a:255). She notes briefly the phenomenon which is explored in relation to the data of this study in Elliott (2010), that examiners showed social engagement with the candidates, their characteristics and their performance, but does not elaborate further. Such engagement does not appear to affect reliability: the one examiner most

responsible for negative affect reactions in Crisp (2008a) was not associated with lower marker agreement, and Crisp implies that the two cases of lowered agreement were due to lack of experience with the specific paper.

Cognitive processes

In terms of cognitive processes of GCSE and A level examining, virtually the only studies that have been published are those done by Cambridge Assessment in the last few years (for example, Crisp, 2008a; Suto & Greatorex, 2008a, 2008b). These used Verbal Protocol Analysis to explore the nature of examiner thinking, in A level geography marking and GCSE Maths and Business Studies respectively, in studies which simulated live marking using a sample drawn from real markers. Suto and Greatorex (2008b) identified five cognitive marking strategies used by examiners in the assessment of short answer question: matching; scanning; evaluating; scrutinising and no response – used when students had made no attempt at the question; they discussed each of these in terms of the dual process model suggested by Kahneman & Frederick (2002), discussed above. They identified difference in the use of these strategies from marker to marker but did not seek to explain these or draw further conclusions from it.

Suto and Greatorex (2008a) developed their findings with a quantitative reanalysis of the data from the (2008b) study, seeking to link different reading behaviours to different types of questions, in particular categorising strategies as being System 1 or System 2, and finding that some Business Studies questions were in general marked using what they termed System 1 strategies, and others System 2. There was a predominance of the evaluating and scrutinising strategies, which Suto and Greatorex identified as System 2. It was suggested that a predominance of System 2 judgements could require a greater element of expertise to enable successful marking, which seems perhaps counterintuitive. If System 1 judgements are resulting in successful marking, it would seem like that these intuitive judgements are based on a high level of individual subject expertise. In Maths, however, a clear distinction could not be drawn between the types of strategies used for different questions; the study concluded that most question parts required a combination of the two types. The authors showed their findings to some experienced senior examiners who validated the strategies and the proportions found through recognition. Nadas & Suto (2009) considered the speed of marking, and found that for markers at all levels of experience, speed of marking increased with practice, a finding which could be used to support increasing automaticity of decisions, shifting cognition from System 2 to System 1.

Crisp (2008a) sought to build on the findings of Suto & Greatorex (2008a; 2008b) in a different subject and age context, looking at reading behaviours, and emotional and personal reactions to scripts. The study compared marking behaviours on short answer and essay questions, and found that different behaviours were exhibited. However, these specific differences are not elucidated in the paper, making it hard to comment on them. Crisp found similar types of reading behaviours to those mentioned in Suto & Greatorex (2008b), above. She also found that over cautious responses where decisions were carefully considered and debated were associated with decreased accuracy of marking decisions, in terms of agreement with the chief examiner's mark. In her discussion Crisp commented on the difficulty of establishing the presence of cognitive heuristics, given that they occur unconsciously. Both these studies acknowledge their limitations in that they used small samples of experienced examiners (six per subject), with only small samples of scripts drawn from earlier live marking processes, and that they addressed only three subjects. The high volume of data from each participant with VPA methodologies necessitates small samples being used, which is problematic for the transferability of the results. Many more studies of this kind are required so that a body of evidence can accumulate, a body to which this thesis is intended to contribute.

Other methodologies than VPA have also been utilised. It is common practice for examiners to annotate scripts while or after coming to a decision on a mark. Crisp and Johnson (2007) undertook a study of the annotation of Business Studies and Maths scripts. Previous studies of annotations (Murphy, 1979; Wilmut, 1984; Massey and Foulkes, 1994; Newton, 1996) have examined whether annotation makes a difference to the levels of agreement between first and second marking of the same script, and have not found any significant effects except where marks were also available. Crisp and Johnson's study involved the analysis of scripts marked and annotated by examiners at home, to investigate how annotation supports the decision-making process. In general they found that questions which had more marks available, because there were fewer questions overall, were significantly more likely to warrant annotation; this may suggest that the cognitive process of decision-making required more support in longer answers. They also found that there was no obvious relationship between marker reliability (as measured by study correlation with live scores) and the use of annotation. Participants were also interviewed about their use of annotation. For the Maths markers, using annotation 'to justify decisions to others appeared to be the most salient purpose' (Crisp & Johnson, 2007:957), with a similar emphasis found in the Business Studies' markers responses. There was also general agreement that it helped

to structure examiners' thinking and the decision-making process. In respect of the proposed research, Crisp and Johnson found

> suggestions that annotations can help to provide a 'visual map' of the quality of answers. This is perhaps especially useful for the purpose of making comparisons – especially between longer texts that possibly exact a great deal of cognitive demand on the marker (2007:959).

The creation of a visual map is aligned with the idea, taken from reading comprehension theories, that a reader builds a mental model of a text, or of different parts of a text (Crisp, 2010b).

That the main purpose of annotation is to 'support the process of the annotator making good judgements' is acknowledged by schools and teachers as well as those within Awarding Bodies and their regulators, where there is also recognition that annotation has a function in accounting for the examiners' thinking to others (Johnson & Shaw, 2010:20). The use of annotation analysis to provide an insight into the examining process is somewhat limited, in that in general the types and amount of annotation are mandated by the mark-scheme – indeed, this was mentioned by various participants in Crisp and Johnson's study (2007). Thus these annotations could not provide a wholly personal account of an individual's thinking, although they can show towards what information the marker's attention has been directed.

There have been attempts to create an overview of the cognitive processes involved. Crisp (2010b) draws on Sanderson (2001), and on her own previously reported study of geography A level examiners (Crisp, 2008a) to suggest a model of the judgement processes involved in examining. The model consists of three main phases, plus epilogue and prologue, and is quoted in its entirety here:

- Prologue – Thoughts before reading begins.
- Phase 1 – Reading and understanding, often with concurrent evaluations of individual points and comments on social perceptions, personal response and task realisation.
- Phase 2 – Evaluation of strengths and weaknesses of the response overall and consideration of how to quantify this evaluation.
- Phase 3 – Mark decision.
- Epilogue – Thoughts after the mark decision. (Crisp, 2010b:7)

With the caveat that each phase can incorporate a number of different behaviours, and that not all phases occur for each item, Crisp found that this model held for different items (ranging from short answer to essay questions), examiners, scripts and examinations (within a single A level geography syllabus). It is a sufficiently general model to be potentially applicable to other subjects, and in the lengthy essay contexts of English and history examinations.

Crisp goes on to elaborate on the stages of the model, in the light of different psychological theories of judgement. All the reading strategies discussed at the beginning of this section, found in Crisp (2008a) and Suto & Greatorex (2008b), are incorporated into Phase 1 of this model. The 'concurrent evaluations' which she includes in Phase 1 are the comments examiners intersperse into their reading behaviour; such comments are the result, Crisp suggests, of rapid comparisons between the 'pattern of associated ideas in the mental model created during reading, and other pre-existing patterns of associated ideas in the examiner's mind' (2010b:9), which she links to the *representativeness* heuristic (as discussed in the previous chapter).

Phase 2, she argues, is a process of 'synthesising or pulling together the important cues in the response and the evaluations made so far' (Crisp, 2010b:13); the implication is that a mental representation of the script is created, so that the final evaluation is based on a short hand version of the script, as with the idea of the visual map created by annotation discussed above. She also suggests that at this point information is grouped and then weighted; where the evaluation does not lead easily into a mark decision, Crisp noted there was evidence of more conscious processing in the verbal protocols, so that an examiner could come to a mark decision. Even in these cases, it was usual for the evaluation in Phase 2 to suggest a range within which the mark should fall; she noted that examiners sometimes reflected on their severity, or explicitly compensated for earlier generosity, for example.

Crisp's model, as she herself notes, is consistent with the use of the heuristics discussed in the previous chapter, and she suggests that it may generalise to other subjects and assessment types. Her article is the only known attempt to synthesise previous research into a model for the judgement and decision-making processes of examiners. It was published after the data collection for this thesis, but does include some similar theoretical underpinnings. I will revisit her model in the final chapter and consider how far it is supported by the data from this study, as well as seeking to establish whether anything can be added to it.

Comparison

It has already been noted in passing that comparison plays a part in the cognitive processes of examiners, but is it a central one? In his influential book *Human Judgment: The Eye of the Beholder*, the psychologist Donald Laming stated unequivocally:

> There is no absolute judgment. All judgments are comparisons of one thing with another (2004:9).

He substantiates this with reference to experiments with human judgement of auditory frequency, visual contrasts, clinical judgements and others, and establishes the principle of relativity firmly throughout the book, although he does not consider it in his chapter on examinations, in which he concentrates on reliability instead. An emphasis on relativity can be confirmed by studies such as that of Crisp (2010a) which compared 'think aloud' protocols of groups of examiners making grading decisions and the same examiners individually marking; in both activities comparison, either to another script or to another response within the same script, occurred regularly, at a frequency of 0.66 and 0.69 instances per script for marking and grading respectively. Crisp (2010b) includes comparison among the cognitive processes of markers without question.

Gill & Bramley (2008) reported a study which used history and physics A level scripts to investigate absolute and relative judgements of examiners, asking them to make three judgements: an absolute judgement of the grade a script was worth; a relative judgement of which of two scripts were better in terms of quality; and an assessment of their confidence in each of the first two judgements. They found that examiners had difficulty in accurately judging the absolute grade-worthiness of scripts, although the history examiners were more able to do so than those of physics. The relative judgements were, however, more accurate than the absolute judgements; interestingly in physics the accuracy of the relative judgement was directly related to the difference in marks between the two scripts, but in history there was no such consistent relationship. Both sets of examiners had more confidence in their relative than their absolute judgements. These findings seem to support Laming's view.

Comparison is already a formal part of the examination system as a whole, with direct comparison of scripts used to produce a rank ordering to check the relative difficulty of different syllabuses in the same subject for the last fifteen years (Pollitt, 2010), and also to check the maintenance of standards over time in the same qualification (Bramley, 2007). Black & Bramley (2008) experimented with the use of rank-ordering, that is, direct comparison of scripts, to cross-validate the results of a grade awarding meeting, and were satisfied that it was an effective validation. Comparison has also been suggested as a more

natural and 'intrinsically more valid' way of making judgements about individual essays (Pollitt, 2004:2), an approach Pollitt has also championed elsewhere (Pollitt & Elliott, 2003; Pollitt, 2010). He has argued that the 'essential purpose' of summative assessment is to provide a rank ordering of candidates, to sort the pool of students, and that therefore direct comparison is both more honest and more useful than attempting to artificially assign scores to essays. His argument is based on the idea that

> in essence we ask our judges to make many micro-judgements and score them so that we can then use simple addition to generate a total score which is used as the macro-judgement required by the Fundamental Purpose (Pollitt, 2004: 5).

The first assumption in this is not necessarily entirely true of English and History A level, although there are some micro-judgements involved, I would suggest 'several' rather than 'many'. However, he has demonstrated the potential of direct comparison as an effective alternative to marking, suggesting that examiners will always compare, even if they have nothing concrete with which to compare a script, reducing them to relying on imagined touchstones (Pollitt, 2010). The findings of Gill & Bramley (2008) discussed above would to some extent support such a move to comparative judgement. Pollitt's method, of multiple comparison of scripts in iterative rounds to create a rank ordering demonstrates a commonality with the 'impressionistic' marking of Britton (1964) in that multiple examiners read each essay, each making a rapid judgement; multiple judgements compensate for speed, and are considered by their proponents to be more reliable. They do have one significant factor to recommend them, which is that such systems move away from the idea of the 'correct' mark, awarded by a single expert in the person of the Principal Examiner.

Mental frameworks

Vaughan, in her paper on holistic rating in EFL, also suggests that

> when papers are read quickly, one after another, as they are in a holistic assessment session, they become, in the rater's mind, one long discourse (1991:121).

She cites the informal comparative statements made by seven of the nine raters as evidence of this. It also resonates with the idea of an overall mental frame of reference into which essays are fitted. The concept of internalised, tacit standards has been accepted by Crisp (2010a) drawing on the psychological terms of Rosch (1978), to describe the 'mental models of likely typical responses' as 'prototypes' (Crisp, 2010a:21).

Wiliam (1998) has termed the use of such mental frameworks 'construct-referenced' assessment; that is, assessment done with reference to the 'construct' or personal understanding of what it means to be a given grade. Marshall (2001), intending to test this, used a qualitative methodology, observing several in-school moderation meetings at which groups of teachers standardised their coursework marking by grading papers supplied by the exam boards. She found that teachers preferred to think in terms of grades rather than marks (something which is possible with internal assessment but not for external examiners, who do not know where the grade boundaries will fall in any given examination session) and quotes a rather telling remark, in which one teacher said 'I instinctively know what it is and adjusted the marks accordingly. This screams D' (Marshall, 2001: 53), a comment which exemplifies the use of constructs of grades by teachers. These constructs were used even though they clashed to some extent with the specific criteria laid down in the mark scheme, with the teachers refusing to be 'bogged down' with the criteria. These quotations are typical of the attitudes which Marshall reports finding in her admittedly small-scale study. Graders interviewed by Baird and Scharaschkin (2002) also suggested that their decisions of grade-worthiness were based on a gut feeling or instinct (five out of ten business studies examiners and two out of nine English). The majority also suggested that their teaching experience helped them to form an opinion as to what a B-grade required; although marking experience had been important, the study was based on the overall grading of a candidate based on all their papers, of which each participant would have experience of marking only one.

Baird explored the process of assigning grade boundaries on the basis of previously marked scripts. Interestingly, she seems to regard the process of making numerical mark judgements as relatively straightforward, because it is 'fairly well specified' in terms of the rubric (2000:91). Instead her study asked examiners of two subjects, English literature and psychology, to decide whether or not 12 scripts, each bearing a numerical mark, were worthy of an E grade or higher, and subsequently to decide what mark constituted the lower boundary, effectively a pass mark. The participants were divided into four groups, matched by severity, who were variously given good exemplars from archive scripts, poor exemplars from archive scripts and no exemplars at all. There was no effect of group on the ability of English examiners to assign scripts accurately or not to grade E, although the use of good exemplars did have an effect on the psychology examiners. Baird concluded that the English literature examiners were 'probably using an internalised notion of standards to carry out a grading categorisation task' although she considered that these internalised standards were somewhat fuzzy and modifiable by the application of different reference material (2000:98).

She also suggested that the psychology examiners might be more used to the idea of an experiment, and had therefore paid more attention to the prompt material. The results of this study both suggest that different subject examiners may utilise different strategies, and also that the ostensible processes may not be the actual processes which occur in examining situations. It did have flaws: the participants were not people who usually went through the process of assigning grade boundaries, so this does not necessarily provide accurate information about the normal process; normally chief examiners are not simply asked to give a mark for the boundaries, but make decisions about whether a number of scripts deserve a grade or not, with the judgements then being aggregated to create the grade boundaries.

However, the reference to mental constructs or frameworks is not always held to be safe; Pollitt uses the word 'imagined' to describe them and asks 'what is the imagined performance that properly embodies a particular verbal descriptor?' (2004:6); it is the nature of the mental framework that it remains unexplicated and internal, so that there can be no way of ensuring the standardisation of the framework. The next section considers two models common in assessment which suggest otherwise.

Guild knowledge and connoisseurship

The concept of a mental framework inside which a specific text or artefact is placed has roots in two models of educational assessment which it is worth considering here. Connoisseurship is more traditionally associated with judgement of art or music, and as such is not well documented in the literature. The general precept, that a novice learns to be expert through extended contact with a master or connoisseur and the example of their behaviour, was captured by Polyani:

> 'the apprentice unconsciously picks up the rules of the art, including those which are not explicitly known to the master.... Connoisseurship... can be communicated only by example, not by precept' (1958:53–4).

Among the types of knowledge communicated in this way Polyani not only included that of wine tasters and art critics, but also that of doctors and other scientists who spend long periods of training in practical skills.

Robbins (2008), as discussed briefly in Chapter 1, argued that connoisseurship was a model which could bring more 'credibility' to the assessment of students' attainment, and in which teachers might play a wider role, as well as being a model under which the 'things we cannot tell' (that is the affective qualities which are desirable to employers and universities

but which are impractical to assess by written examination) might be judged. It is a model which would be compelling to those who might consider themselves to be connoisseurs, and to those opposed to the 'tests and targets' mentality which Robbins criticises; he uses it primarily as a theory by which teacher assessment might gain authority. There has been no clear attempt to explore the role that connoisseurship might play in formal examination systems, although the idea of a novice gaining expertise by an almost osmotic process through carrying out judgements together with an expert does bear a certain resemblance to the process of the training meeting and the team-based exercises.

The metaphor of the novice and master is a powerful one, which has also gained credence in the form of 'guild knowledge', as coined by Sadler (1989) and considered more recently by Marshall (2011). Here the indoctrination of the teacher into the 'guild' gives them access to a shared knowledge; Sadler discusses it in the context of formative assessment which, by

> 'providing guided but direct and authentic evaluative experience for students enables them to develop *their* evaluative knowledge, thereby bringing them within the guild of people who are able to determine quality using multiple criteria' (1989: 135; italics in the original).

In the same way teachers gradually gain guild knowledge by shared assessment and discussion, and through a set of shared values of what it means to be 'good' at the subject. Marshall, while adding the critique that this does not account for potential differences in points of view about a subject (2011), links this to what Britton (1964) termed 'impressionistic' marking (as discussed under 'Holistic versus analytic marking' above), in that both involve a shared knowledge which it is not necessarily possible for the teacher or guild member to articulate. Wiliam, drawing on an alternative terminology but still echoing these concepts, argues that teacher assessments of this kind 'are construct-referenced assessments, validated by the extent to which the community of practice agrees that the student's work has reached a particular implicit standard' (1998:7). The concept of the construct-referenced, rather than criteria-referenced, assessment, in which scripts are held up against a mental construct of 'what an A is', for example, fits well with the idea of the mental representations explored in the previous section, and also with the use of comparison as a primary cognitive decision-making tool, as considered above.

Both connoisseurship and guild knowledge could be used to explain studies in which there is an apparent ability of groups of teachers to make similar qualitative judgements

without resorting to highly specified mark schemes or bodies of criteria; indeed they have arisen out of such apparently inexplicable data. The mental representation or framework within which any given essay or script is finally situated is an idea which is well established in the literature, in both empirical and theoretical works; the data within this study will also be examined for evidence of such a framework.

The training meetings

The standardisation or training meetings form an important part of the process by which examiners learn the required standard of marking, familiarise themselves with the mark scheme and come to understand the mandated marking behaviours. Their centrality to the decision-making processes of examiners may be exemplified by the fact that one of the 'key features' used by inexperienced markers in Crisp (2010a) was one which the team leader had mentioned. Sfard (1998) suggests two metaphors of learning: acquisition and participation. The training meetings draw on the metaphor of participation as a particular technique by which to indoctrinate the examiners into the guild, and to refresh the knowledge of the experienced examiner, as groups of examiners, of different levels of experience, participate in the marking process with the expert, or senior examiner.

If there are few studies considering the cognition of examiners, there are fewer still which deal with the standardisation process, and in particular the training meeting which has remained a closed process. There are two detailed studies which have looked at the live standardising meetings; Daugherty (1988) and Sanderson (2001). The study of Daugherty (1988) was published as a report by the Secondary Examinations Council, so is not widely available. It sought to establish a description of the process by which candidates were awarded grades in two Geography O level examinations in the years 1984 and 1985, and drew on observations of the question-setting meetings, the standardisation meetings and the grade awarding meeting (at which the grade boundaries are decided). Despite the time which has passed since this research it is worth considering in some detail, since it provides a more complete account than any of the journal articles available, and is almost unique in its examination of the live process of examination grading. Its scope did not extend to the individual decision-making of examiners. Sanderson (2001) draws on eleven years of 'participant observation' as a Sociology examiner at GCSE and A level, as well as the observation of the pre-standardising and standardising meetings for one examination each of Law and Sociology for the main data gathering phase. His study gives a descriptive outline of the training meetings rather than analysing them in detail.

Daugherty establishes three key features of the 'examiners' conference' (which is referred to in this study as the standardisation or training meeting): the discussion is focused on actual scripts; all possibilities are considered; and discussion continues until there is an agreed marking scheme (1988:17). The examiners work on a draft marking scheme, in conjunction with scripts of actual answers, and ensure that all possible answers to a question have been considered; in one of the two meetings the aim was for an exhaustive list, while in the other examiners were encouraged to use their own professional judgement. Sanderson (2001) also reports the alteration of mark schemes during pre-standardisation meetings (in either law or sociology; he did not distinguish between the two in his data). This differs from the current practice in English and history, in which the mark scheme is set before the examiners meet, and there is no consideration of other possible correct answers; this is due to the difference in subject (as confirmed by current geography examiners), and the nature of short answer questions, rather than to changing practice over time. To some extent history examiners also have to consider different possible answers, but the assessment model is not one of short answer factual recall but of weaving of factual knowledge into argument for an essay; that is, the model which is considered in this thesis.

Daugherty considers the chief examiners to have 'retained the high ground of the decision-making process' (1988:24) although they were challenged at times by the other examiners, and indeed in one of the meetings, the smaller one, he describes the style of the decision-making as 'consensus'. In one of the meetings he describes, on a small number of occasions, the group of examiners settled on a decision by a majority vote. The process which Daugherty describes is one in which there are many inputs to the overall process of judgement; 'each participant in the examining process brings to it a personal perception, based on experience, which will influence their judgement' (1988:46). It is clear that for this syllabus at that time, the standardisation process could be conceptualised as being one of making decisions as a group, with negotiation where necessary and potentially using the 'contest' model of decision-making as conceptualised by Christie & Forrest (1981). How far this conceptualisation holds true for the data from the current study will be examined in Chapter 4 below.

Sanderson's main concern was to investigate how the standardising meeting creates a 'community of practice', which is his chosen framework. Drawing on the psychology of judgement literature on heuristics, he proposed a 'common cultural heuristic' inculcated by means of the standardising meeting, in line with Wolf's (1993) view that socialisation, rather than detailed specification of criteria, leads to reliability. In his data, he notes variation in

how much challenge was allowed to the Principal Examiner's view, but adds the caveat that in subjects where there are very large numbers of candidates, and therefore examiners, 'transmission of judgement tends to be rigidly hierarchical' (2001:103). He cites an interview with a Senior Examiner for English in support.

Other studies which have been reported, which are few, have focused on the accuracy of marking subsequent to variations in the training process. Raikes, Fidler & Gill (2010) conducted an experiment with 24 markers of Psychology A level, experienced and new, in which they replaced the face-to-face meeting element of standardisation with extra written material, including a mark scheme rationale and a written explanation of the marks awarded to standardisation scripts. They concluded that while standardisation improves the accuracy for both new and experienced examiners on both short answer and structured factual essays, the benefits of including a face-to-face meeting are variable, small and questionable. They do attach the caveats that the essays they used were highly structured factual essays which were marked against a prescriptive mark scheme and that their 'findings might not be replicated with less constrained essays and marking' (2010:26). Greatorex & Bell (2008) also used reliability as a measure of success for their three standardising interventions for biology markers, being a traditional standardising meeting, personalised feedback and pre-written feedback (the first two being the usual approach). Most combinations of the three types improved reliability, although they did find that standardising meetings alone did not have much effect for experienced examiners. Both of these studies, as Raikes, Fidler & Gill reflect, were simulation studies, and the participants knew this. A simulation study not only does not reflect the real pressures of the examination marking process, it also means that the participants have to be even more motivated and self-selected than in research which takes place during natural events, which may have an effect on their ability to mark or their response to training.

Conclusions

The literature on examining is a disparate one. What can be seen is that there is a vast array of potentially relevant factors and processes which have been identified by studies looking at a variety of subjects and answer types. There is a dearth of recent research on essay-based assessment, particularly in the context of end-of-school examinations. With the notable exception of Crisp (2010a) there has been no attempt to synthesise research to provide an overall consideration of the cognitive mechanisms and heuristics which enable examiners to make sufficiently reliable marking decisions without using either more time,

memory or attention than they could have available. Drawing on this literature, the next chapter will establish the research questions which have driven this study, and outline the methods used to address them.

Chapter 3: Methodology: how to look inside the black box

Overview

Research into cognitive processes is not straightforward. The mind has been characterised as a 'black box' system: it is possible to observe stimuli and responses, but not what happens between the two, inside that black box. Recognition of the full range of cognitive processes which occur in any given activity is not available to the conscious mind of the individual. This study follows a qualitative design, based on an initial stage to gather rich data from a small sample. The first stage uses analysis of recordings of examiner training meetings and of 'think aloud' or Verbal Protocol Analysis recordings to infer and deduce the cognitive processes which are taking place. The second stage took the initial findings and presented them to a larger sample of examiners, in the form of a questionnaire, to gain a measure of validation and to gather further data about how examiners view the process.

Research Questions

The research was driven by the following over-arching question:

How do examiners make their decisions when assigning marks to essay scripts?

Examiners do not operate in a completely isolated world. Their decisions are taken in the context of a variety of tools which are provided to help them mark according to the expected process; not least is the training meeting which familiarises them with the paper and the mark scheme. Without consideration of the training meeting, it would be impossible to trace whether elements of their decision-making were entirely individual. In addition, the training meetings have not been the subject of research in the past, so are of interest in and of themselves. Beyond that, the cognitive process of examiners when making judgements of essay scripts can be explored by close analysis of transcripts of their speech, to deduce the behaviours which they engage in, the characteristics of which they take account and the methods by which they reach a decision. With that in mind, the following sub-questions were identified:

1) *What decision-making behaviours do examiners exhibit?*
2) *What training strategies are demonstrated by senior examiners and how do they relate to the decision-making behaviours exhibited?*
3) *What foci of examiners' attention during reading can be identified?*
4) *What tools do they use to make their decision and what is the role of the mark scheme?*

In order to explore these questions, data were gathered during the training meeting from groups of examiners, including senior markers, and afterwards, from individual examiners during the course of their normal marking. Standardising meetings allow the modelling of desired behaviour and thought processes by senior examiners to their juniors; in order to do this they must make their thinking and behaviours explicit. Reciprocally, all examiners have to make the thoughts behind their decisions clear, as far as possible, to their seniors in order to justify their decisions and to present themselves as competent in judgement.

1) What decision-making behaviours do examiners exhibit?

Decision-making behaviours can be both physical actions and cognitive strategies. Both, to some extent, can be seen through the speech that examiners produce. Most of the behaviours that require physical actions are likely to relate to the tools which examiners use, and are therefore relevant to sub-question four, below. However, the cognitive behaviours which they exhibit, such as those which relate to anchoring and adjustment, for example, are identifiable in their speech through deductive reasoning, even if the examiners themselves are not wholly conscious of the cognitive processes in which they are engaging. The recordings of both the training meetings and the 'think aloud' sessions generate data during normal decision-making and during training activities designed to practise decision-making, which is then examined for evidence of the cognitive practices of the examiners. This question is the one which relates most directly to the theory of heuristics and biases examined in Chapter 1, and to the principle of comparison which is discussed in Chapter 2, and in relation to the data in Chapters 8 and 7 respectively.

2) What training strategies are demonstrated by senior examiners and how do they relate to the decision-making behaviours exhibited?

This question relates to the organisation of the standardising meetings, and the ways in which senior examiners talk to their examining teams, as opposed to the organised training activities relating to marking standardising scripts. The recording of the standardising meetings provide rich data for exploring the ways in which senior examiners try to inculcate appropriate marking and decision-making behaviours, and the interpersonal strategies in which they engage to adjust examiners' behaviour to the correct model. Training is carried out by principal examiners and team leaders; the verbal data show both how they make decisions and how they try to influence the decision-making of others. The responses of the

examiners being trained may show the specific effect, if any, of the individual trainers. It is also possible to link the traits exhibited in the individual 'think aloud' recordings of some examiners to the training which they received.

3) What foci of examiners' attention during reading can be identified?

On the basis of the literature discussed in Chapter 1, it is clear that attention is limited and therefore anything which forms its focus during the marking period is in some way significant, and potentially a determiner in the decision to be made. Anything which is vocalised during either a training meeting or a 'think aloud' session has, by definition, been the focus of attention at least at the time of vocalisation. Further literature on the use of 'think aloud' or Verbal Protocol Analysis to identify the items to which an individual pays attention is discussed below. Characteristics and features of scripts on which a judgement is based must be among those which attract the attention of examiners; similarly the attention which examiners pay to the mark scheme, indicative material, sample scripts, or other items during marking will illustrate the mechanisms of their decision-making. During the training meeting senior examiners provide on-going commentary on scripts, illustrating explicitly the passing focus of their attention.

4) What tools do they use to make their decision and what is the role of the mark scheme?

It is the expectation of the Awarding Bodies that the mark scheme will be the most significant tool in examiners' decision-making, although they provide a number of other potential tools, including the sample scripts. The tools which examiners mention when they are reaching judgements on essays should enable the identification of those which are of most use, and whether they are drawn exclusively from the expected resources or from outside those provided by the Awarding Body. The mark scheme itself deserves a more thorough examination, sitting as it does at the centre of the official process; its role will be identified by noting when and how examiners refer to it, both during the standardising meeting and in their own individual marking.

The findings chapters have not been organised according to the research questions, although there are certain groupings. The first two findings chapters are essentially concerned with sub-questions 2 and 3; they consider the training meetings in detail, and the commentary on scripts within the meeting. Material relevant to sub-question 3 is found in other chapters, however; similarly sub-question 4 is considered in several chapters, although

the third findings chapter deals with the role of the mark scheme in some detail. Chapters 7 and 8 focus on the cognitive decision-making behaviours of sub-question 1, although these also appear in passing in other chapters. The final findings chapter reports the results of the questionnaire, which speaks to all the research questions.

Methods

Phase One

Sampling

Examiners were recruited through a host organisation, one of the four main Awarding Bodies in England and Wales, for which the pseudonym 'Exam Board' has been adopted. The examiners who take part in marking during January are experienced and 'good' markers, as chosen by Exam Board from the larger pool of summer markers. Four modules in total, two from A level History and two from A level English Language and Literature were chosen to form the basis of the study. This represents all the modules which are examined in those subjects in January. A general invitation to participate in the recording of the standardising meeting and the 'think aloud' study was issued to all examiners on each module. Informed consent was sought from all the Principal Examiners and Team Leaders for the four modules which form the basis of the study. Consent was also sought from all the examiners for the online modules, all the participants in the live English meeting, and from all examiners in a single team for the History meeting. No examiner refused to allow their participation in the meetings to be recorded.

A smaller selection of examiners volunteered to take part in the Verbal Protocol Analysis. The sample was self-selected, by their willingness to participate. An inducement was offered to encourage participation, a decision which is discussed below under 'Ethical Considerations'. Very few examiners volunteered to participate in this part of the research, and many fewer than I hoped. A sample of five agreed to take part, spread across all four modules; only three completed the recordings and returned them, two examiners from History 1 (Parrot and Caspar), and one from English 2 (Lupin). Lupin is also identifiable in the training meeting; unfortunately the break-out meetings chosen on the basis of 'think aloud' participants were those of the two who did not return their recordings. The 'think aloud' data is taken from a much smaller sample than would have been ideal, but it provides complementary data for the recordings of the training meetings.

The sample thus consists of: all Principal Examiners and Team Leaders, and *circa* 40 examiners for the recording of the standardising meetings, with a smaller sample of three providing the 'think aloud' data.

Data Collection

a) meetings

Five meetings were recorded: the online training meetings for History 1 and English 1; the live pre-standardisation and standardisation meetings for History 2 and the live standardisation meeting for English 2. Each meeting was between four and six hours long, providing a total of approximately 25 hours recorded data. A digital voice recorder was used to make the recordings; these were supplemented with extensive verbatim note-taking, in the live meetings, which protected against the recordings being unintelligible as a result of background noise. The majority of the speech in any given meeting is from the senior examiners: the principal examiners conducting the meeting, or the team leaders who are training their individual teams. The number of examiners who speak during a meeting varies between modules; even on the Voice Over Internet Protocol meetings a minimum of six normal examiners will also contribute. (The exception being English 2, which had five participants in total, two normal examiners, the Principal Examiner, the Chair of Examiners and one other senior examiner).

These individuals nominally work in teams, which are temporary groupings based around a Team Leader in the modules with a larger candidate entry, or around the Principal Examiner him or herself in one with a smaller entry. The teams consist of up to 12 examiners. The recording was taken of the plenary meeting in all modules, and then of one team in each of the two history modules, during the break-out session. Although they nominally work in teams, the examiners are not engaging in group decision-making; rather, they are reaching individual decisions, although they may be influenced by the other members of the group. Indeed, the tendency of individuals within a group to conform to what they perceive as the norm (Bettenhausen & Murnighan, 1985) could be favourable as far as the Awarding Bodies are concerned, providing the dominant group member is the senior examiner.

The meetings are of interest in their own right. No prior research has been carried out into either the pre-standardising or the standardising meetings as they normally function[1]: they are closed systems through which key decisions are made which have a major impact on the outcome of the examination session. It is at these meetings that the senior examiners rationalise their decisions and verbalise them in order to train the normal examiners. In addition, without knowledge and understanding of what happens at the specific standardisation meeting which examiners have attended, it would be impossible to draw reliable inferences from the 'think aloud' protocols recorded by them. The way in which they address and interpret the rubric is shaped and defined by these meetings, and decisions which show lack of reliance on the rubric may be coming from their *own* rules, or from 'rules of thumb' supplied by peers or supervisors at the training meeting. Knowledge of these meetings is essential to distinguish between the two.

The principles behind the recording of these meetings draw on the principles of discourse analysis, in the wider application of the term. This acknowledges that although cognitive processes are the primary concern of this study, examining is a social activity that takes place in a community of examiners, drawn from a teaching context, and where the social constructs affect interactions; nevertheless these interactions also illustrate underlying mental representations. In particular one of the underlying principles of this study is that there is a limit to the mind's attention capacity; discourse analysis provides a method by which the foci of attention can be identified, and thus the concerns which guide cognition.

The recordings of the training meetings bear a strong similarity to the 'think aloud' data, in that they represent a group of people explaining their thought processes to each other, albeit with a social function, and the adjunct differences, rather than a single individual voicing their thought processes to themselves. Some of the literature concerning examiners' decision-making supports the standardising meeting as a valid source of data, as research such as that described by Hamp-Lyons (1991) effectively constructed a training meeting as a method for uncovering how raters think about essays. The standardising meeting provides a forum in which examiners must convey their judgements about essays; from seniors to juniors this communication is a model of their perception of ideal practice,

[1] Greatorex & Bell (2008) and Raikes, Fidler & Gill (2010) both investigated the impact of *altering* the format of the training meeting, but there is no published research on the normal content or process of these meetings.

and from juniors to seniors it exemplifies and justifies their reasoning behind the judgements which they make.

b) individual examiners

The dominant methodology used in studies of examiner thinking is Verbal Protocol Analysis (VPA) which is also known as the 'think aloud' method. VPA is based on the assumption that 'an individual's verbalisations may be seen to be an accurate record of information that is (or has been) attended to as a particular task is (or has been) carried out' (Green, 1998:1-2). It differs from early attempts at introspectionism, where participants attempted to report their own cognitive processes, in that it accepts that this is impossible and instead attempts to make inferences about those processes from the information which has been stored in the individual's short-term memory. In a review of the literature on tracing mental processes Ericsson and Simon (1980) examined more than fifty studies using variations on the 'think aloud' method and concluded that where verbalisation is 'concurrent' (information is verbalised at the same time as the participant is attending to it), is based on verbal stimuli, where filtering of information is not required and there is no extra requirement, for example to create verbal descriptions of their motor activities, there is the least possibility of interference either with the base task or with the information which is being verbalised.

Graesser *et al.* approved the use of 'think aloud' protocols as providing "a very rich source of data for discovering possible comprehension strategies and for testing detailed claims about the representations that enter the reader's consciousness" while warning that "protocols do not reliably tap unconscious comprehension processes" (1997:166).

Crisp (2008b) validated the use of 'think aloud' techniques in the context of studies of examiner thinking, using a design where a sample of markers of Geography A level marked a selection of scripts at home and then at a one-to-one meeting with a researcher, marked a further sample, some silently and some while 'thinking aloud'. At this meeting each participant marked two scripts already used during their home marking phase, one silently and one aloud. A total of six scripts were used in this way, so each participant marked two of six common scripts, twice. All examiners were found to be slightly more severe during the study than during live marking. The mark difference between home and silent at meeting marking and the mark difference between home and 'think aloud' at meeting marking were compared for each examiner and found not to be significantly different. Further analysis of

the overall study indicated that thinking aloud while remarking scripts did not cause 'greater than normal variation in mark to re-mark differences' (Crisp, 2008b:8).

Both Crisp (2008b) and Suto and Greatorex (2006) also asked examiners for their views on whether thinking aloud had affected their marking. Some examiners in both studies identified a slowing in their thought processes due to the 'think aloud' protocol, although some felt that this had been beneficial and had led to more thoughtful consideration of the script and more accurate marking, while others found that this had distracted them from the flow of the essay. Crisp (2008b) finishes by recommending that 'think aloud' protocols are not used for the marking of live scripts in case of an effect on the candidate's mark, despite their evidence to the contrary. Neither, however, found any evidence that the 'think aloud' technique altered the psychological processes underpinning the examining decisions. Ericsson and Simon firmly state that where the conditions described above are met, there will be no change in the course, structure, or speed of cognitive processes (1993).

Three examiners provided 'think aloud' data: two provided three hours each, spanning the examination marking period (Lupin and Parrot); the third provided an hour's recording from the beginning of the marking period (Caspar). Each examiner used a computer based audio recording programme to record themselves during live examination marking, which all takes place on-screen over a secure internet connection for A levels in the host organisation. They were given the following instructions:

> I would like you to literally 'think aloud' while you mark for one hour. You should do what you normally do when marking, but speak aloud everything that goes through your head, whether it seems relevant to you or not. It would be helpful if you can also describe what you're doing, for example if you make an annotation about an answer, or flick back to look at the mark scheme. Please try to vocalise as much as possible. I know that it is difficult but the essential thing is that you should remember to keep talking.

As a result of the pilot the decision was taken to ask examiners to record themselves remotely, without the presence of a researcher. A clear theme emerged from pilot participants' comments that they would be less inhibited 'thinking aloud' at home, without a question prompting about location; to a lesser extent they thought that having someone in the room with them acted as an inhibitor. Nor was the added information gained by

observation during the pilot recordings considered to be worthwhile enough to justify the added difficulty and intrusion of recording the examiners in person.

Data Analysis

The data from both the 'think aloud' sessions and the training meetings are treated in a manner which draws on Verbal Protocol Analysis for its analytic philosophy. It differs from a traditional discourse analysis approach which places emphasis on linguistic content and structure. Although similar processes are applied, in Verbal Protocol Analysis 'inferences are actually made about the cognitive processes that produced the verbalisation' (Green 1998:1); the focus is on the thought behind the utterance, rather than on the socio-linguistic processes which produce it. The exception to this is Chapter 4, which draws more strongly on the field of traditional discourse analysis to explore the linguistic structures which are used by trainers and how these relate to the training strategies. The linguistic structures of the essays, and their pragmatic relation to the examiner, is a subject for analysis, but in terms of the examiners' responses to, and verbalisation of, those structures.

All the recordings were transcribed, therefore, and treated in the same way. Analysis proceeded in two ways: bottom up, using repeated readings to gain an understanding of the themes which were emerging from the data; and top down, using the theoretical framework which had been pre-established to look for relevant evidence. In this way I attempted to avoid both the potential pitfalls: of finding evidence that confirmed only pre-existing ideas, and of having so few guiding principles that no coherent findings emerged. Initially a sense of the overall process of each recording was sought, before more specific coding and analysis began. Some initial codes were identified and a coding sheet produced, which was updated and significantly altered as time went on.

All scripts were annotated to begin outlining interesting patterns. To exemplify, one feature which emerged early on from the 'bottom-up' analysis was comparison. All episodes of comparison were coded; sub-codes were then established to differentiate, for example, comparison between the script and a sample script or between the script and the mark scheme. An example of top-down analysis would be the identification of evidence for the use of different heuristics by examiners; this also incorporated elements which emerged from the text as rules of thumb which were suggested as short cuts by the senior examiners were compared to the heuristic categories and sorted or identified as stemming from elsewhere.

In contrast, attempts to code for the linguistic theories of pragmatics in the detailed way outlined in the discussion of Graesser *et al*. (1997) in Chapter 1 above, were not fruitful. Instead, its direct use was abandoned in favour of the emergent category of 'Quality of Written Communication', which is not precisely equivalent, but which was much more present in the data. The theoretical background of pragmatics is used to understand and support this theme, which is a specific aspect of the Awarding Bodies' systems of examination.

Broad themes were established, and the transcripts were re-read for relevant instances of those themes. Then those instances were compared and categorised within the theme. A measure of confirmation was sought by presenting data and preliminary analysis to a supervisor and fellow research students: agreement was reached that the interpretation being placed on the data was reasonable, and that coding was being applied appropriately.

The excerpts of data which appear in the body of the thesis to illustrate points were chosen for two characteristics. Firstly, they may be the most typical examples of the characteristic under discussion, or they may have been chosen for clarity (of meaning: it is the nature of talk that when utterances are de-contextualised they can become almost unintelligible to an external party). Words which are the candidate's, read from the script, are indicated by italics and the markers *« »*. Quotations from the data are always attributed to the individuals concerned; sometimes abbreviations are used, e.g. 'PE1, H1' for 'Principal Examiner 1, History 1.' (A full list of abbreviations can be found on page 4.) The examiners are numbered in the transcripts of meetings; in English 2, where there were only two examiners who were not senior, they are designated 'Examiner' and 'Examiner*' (the latter is Lupin).

Phase Two

Methods

A questionnaire (see Appendix) was constructed using the findings from the first stage of data collection, divided into three sections: statements taken from the phase one data with which participants were asked to agree or disagree; open and closed questions about the individual examiners and their experience of marking; and a series of Likert-type rating questions with summative statements drawn from the conclusions reached by analysis of the first phase.

Section one comprised 34 randomly ordered statements quoted or paraphrased directly from the transcriptions of data (with one exception), expressed in the first person.

Participants were asked to consider whether the statements were something that they recognised from their own thoughts while marking examinations, with a simple yes/no answer. The statements are spread across the themes of: comparison (7 items); quality of written communication (3 items); Heuristics (9 items, divided between anchoring and adjustment, representativeness and unofficial heuristics); Affect (4 items); Self-adjustment (2 items); deduction (4 items); and the mark scheme and marking behaviour (5 items). One statement was expected to provoke a 'no' response (this is the statement which is not taken from the transcripts); one pair is a contrast pair, with a participant expected to assent to only one of the two. A breakdown of the statements by theme may also be found in the Appendix.

Section two consisted of mainly open questions. These did not speak directly to the topic of decision-making, but aimed to gather more data on some topics which were raised by examiners outside the recording, and to provide further contextual information to support or challenge the conclusions reached on the basis of the first phase data. This section also contained some closed questions to ascertain the subject for which the participant was an examiner, the format of the training meeting and how long they had marked.

Section three presented those conclusions directly to the participants. Six groups of three statements each were presented thematically, on the topics of informal heuristics, anchoring and adjustment, deduction, quality of written communication, representativeness and comparison. Each was accompanied by a four-point Likert scale, ranging from 'strongly agree' to 'strongly disagree'. No neutral option was offered, although participants did have the option of ignoring the question. (Six participants took the option of passing on a question, one of whom ignored two statements. Five of these missing answers referred to questions in the third section.) The statements were expressed in the third person, unlike those in section one, and presented conclusions in a fairly straightforward way. They did run the risk of provoking what might be termed the 'social acceptability bias' (Robson, 2002:233) in that they suggest that what might be thought of as bad marking behaviours are the norm.

The questionnaire was delivered through the medium of an internet web survey site. This enabled swift collection of responses, with complete anonymity guaranteed to the participants.

Sampling

Recruitment for the second phase was by invitation relayed through all four of the main Awarding Bodies in England and Wales, to all examiners marking modules in A level English (Literature, Language or Language and Literature) or History qualifications during the January session of 2011, after the conclusion of that session. Participation was entirely voluntary, with no inducement offered. The sample consists of 45 participants who completed all sections, split almost equally between history and English examiners (23 and 22, respectively); a further single participant fully completed the first section but not the demographic data or the third general section. The number of years' experience as A level examiners which participants could claim was impressive ($\bar{x}$ =19, S_N = 11.6); they represented a wide range, from 2 years to 43 years. A small number of participants identified themselves as team leaders or principal examiners in their answers to the open questions. The modules which all participants were currently marking were standardised using face-to-face meetings.

Data Analysis

Responses to the closed questions from the questionnaire were entered into a database (SPSS) for analysis. Simple count analyses were used for the yes/no questions in the first section of the questionnaire, and similarly to gauge the support offered for the statements in the final section. Independent samples t-tests were used to establish whether there was a statistically significant difference between the responses of examiners of different subjects, or of different levels of experience. Cross-tabulation analysis was used to examine any correlation between patterns of response. The open question responses were analysed question by question for patterns of responses. Simple inductive coding was applied to group the responses thematically. The coded responses were also compared across questions.

Ethical Concerns

In the context of the examination system the prime ethical responsibility has to be to the candidates whose futures depend on the outcome of the examination at hand. Previous research has depended on simulation studies, to avoid any question of altering candidates' outcomes. This has implications for the validity of the study, in that it can always be argued that results from a simulated task may be unrepresentative simply by virtue of being part of a simulated situation, rather than a live situation. However, as shown in the discussion of the

method for data collection in Phase One, above, the evidence from the literature shows that 'think aloud' protocols may slow processes but not change the outcome. Exam Board has its own strict protocols in place to monitor the quality of marking, which all continued to be observed during the study.

The question of recording the training meeting required consideration in terms of the possibility of affecting the fairness of the examining procedure. If the presence of recording equipment disturbed the participants in the meeting, it might have interfered with their learning process and jeopardised their accurate marking, therefore damaging both them and the candidates whose scripts they mark. However, the fact that the training meetings take up the best part of a day, and that nerves would surely be overcome with rapidity, as the case proved, was considered to mitigate the potential danger. Participants were most eager to ensure that I was independent and not being paid by their Awarding Body, after which they were completely unafraid of my presence. Additionally all the participants in the team being recorded gave their consent to taking part, so anyone truly uncomfortable with being recorded would have excluded themselves. Recording equipment was as physically un-intrusive as possible.

The ethical consequences of being 'board-sponsored'

The only way of gaining access to the data necessary for answering the research questions was to go through an awarding body. The danger then arose that those approached to participate might feel compelled to take part because of their status in relation to the awarding body, i.e. that of employee. Equally they might fear that details of their marking performance, which might compromise their being offered future employment, would be revealed to the awarding body, which might alter their behaviour or inhibit them from speaking honestly about their processes. Ethically this is problematic because of the discomfort that individuals might experience over participation. The information which was provided to them was key in dispelling this. During the recruitment procedures participants were given an undertaking that what each of them said would not be repeated to the awarding body in any way attributable to the individual. Such an undertaking was supported by the representatives of Exam Board with whom I met during the negotiations for access. Clearly there was no feeling of compulsion, given the small number of participants who volunteered for the 'think aloud' task.

Since recruitment had to take place via the awarding body, it might have been possible for them to deduce the identity of the group of individuals from whom the data

comes. Given that awarding bodies have their own methods for ensuring the quality of marking, and that the aim of this study was to answer the research questions using data gathered from a range of methods and participants, it seemed unlikely that they would go to such lengths, and no evidence was presented to the contrary.

The ethics of incentives

After discussion with Exam Board I concluded that it would be necessary to offer a material incentive to participants in the 'think aloud' study. The population from which the sample was drawn was composed of individuals who mark examinations as an extra source of income. Marking is time consuming, and participation in a study which required the expenditure of extra time and effort seemed to justify some reward. There was no prospect of offering a 'professional gain' as an incentive, since the outcomes of the research are not of use specifically to the individual examiners, only to the Awarding Body as a whole. The incentive offered was a chance to win one of two gift vouchers for books, as a token acknowledgement of the effort involved.

BERA guidelines on the use of incentives state:

> Researchers' use of incentives to encourage participation must be commensurate with good sense and must avoid choices which in themselves have undesirable effects (e.g. the health aspects of offering cigarettes to young offenders or sweets to school-children). They must also acknowledge that the use of incentives in the design and reporting of the research may be problematic: for example where their use has the potential to create a bias in sampling or in participant responses (BERA, 2004:8).

In this case the first concern is not applicable; the second concern is that the use of incentives has the potential to create bias in sampling. In one way this was not of concern as demographic characteristics such as socio-economic status are not considered as factors within the research, and in any case the population is of a relatively homogenous nature, since examiners are typically drawn from the teaching profession. Secondly, in this case I would argue that withholding incentives would have created a greater bias in the sample than using them: this is an activity for which individuals are usually paid. Those who agree to increase their workload without increasing their payment are likely to be those with higher economic standing. However, offering payment would not lead those people to exclude

themselves from the study. Previous studies of this nature have used financial incentives; Suto & Greatorex (2008b), for example, paid their participants.

Most of the discussion around the ethics of participant payment has taken place in the field of medical ethics, since participants in medical trials may be vulnerable by virtue of the condition which makes them eligible for the trial, or there may be the potential for healthy volunteers to come to physical harm by participation. However, despite the different stakes involved in this particular study, much of the debate is still relevant. Possible ethical concerns which have been raised by commentators on the issue of payment for participation in research, were summarised as follows by Wertheimer and Miller (2007). Payment for research may:

- Compromise integrity of research, because participants withhold important information
- Wrongly commodify a practice which should be based on altruism
- Create injustice if some groups are more likely to respond to financial incentives than others
- Be coercive and thus compromise voluntary consent

The first difficulty did not apply since recruitment did not depend on the information which participants provided, nor was there any 'seeding out' once the research had begun. The question of injustice is related to that of bias, dealt with above. 'Injustice', however, implies the possibility of damage resulting from participation in the research, or that it is vulnerable groups who may be exploited because of payment. I have already stated above that I do not believe the latter to be the case in this study, and it seems to be highly unlikely that any damage could have accrued to participants as a result of the research.

The second and fourth arguments are more complex. The question of whether offering payment to participants constitutes coercion is one which has been extensively debated in the context of medical research ethics. It is self-evident that coercion 'invalidates consent' (Wertheimer & Miller, 2007:390), and that research that utilised coercion would therefore be ethically dubious. Whether or not payment can be said to be coercion depends on the definition of coercion which one uses. Wertheimer and Miller cite one view of coercion as being an offer to which there is 'no reasonable alternative' to accepting. This is the situation in which a financial offer would be so large as to overcome other considerations which would lead an individual to reject the offer to participate. In this study the payment could in no way have been large enough, or certain enough, to distort the judgement. Alternatively, coercion can be characterised as telling an individual that unless

they agree to do something, you will cause them harm, by action or by omission of action. Wertheimer and Miller eventually conclude that the offer of financial payment cannot coerce.

This leaves the argument that offering payment taints the altruistic act of volunteering to participate in research. Altruism is not a characteristic which can be forced on the participant, and nor, in my view, does offering compensation for the time and effort which they have expended, lessen the value of their act. It is simply offering an appropriate recompense.

Confidentiality

I have considered above the issues relating to confidentiality of the participants' data with regard to the host awarding body. There are two further issues relating to confidentiality: dealing with scripts which contain the work of individual teenagers, and dealing with data which have a commercial value to the host body, in working within their business. It was agreed with Exam Board that the contents of individual scripts and the mark scheme would remain confidential, until they were released into the public domain by Exam Board, which they were six months after the main phase of data collection. No names of candidates or details by which candidates might be identified are used in this dissertation, nor will they be in any publication arising from the study.

In terms of the confidentiality of commercial data, anything which I learned in the course of my research which was not covered by the remit of the study has been kept strictly confidential.

Chapter 4: Characterising the Training Meetings

"For legislators make the citizens good by forming habits in them, and this is the wish of every legislator, and those who do not effect it miss their mark." Aristotle

The standardising meetings which featured in this study were all very different in character, despite their common purpose, and their substantial similarities in form and structure. For contextual understanding, and because between them they cover the range of standardising meetings that take place in all subjects across awarding bodies, the main characteristics of each meeting will be outlined below. The remainder of the chapter will explore the ways in which the training meetings are conducted, and the strategies employed to train examiners. Although every trainer's style is different, there are common approaches across the whole, some of which can be seen to be deliberate, and others perhaps instinctive. The training meetings mandate two sets of behaviours: marking behaviours, such as when or at what rate to mark; and decision-making behaviours, associated with the kinds of processes that are outlined throughout this thesis. However, they also seem to inculcate certain attitudes; in particular, a sense of responsibility towards the candidates, and an emphasis on fairness and justice. These attitudes are linked, explicitly and indirectly, with the desired marking behaviours, and are clearly associated by senior examiners with appropriate and rational application of the process.

The system of standardising meetings is to some extent aligned with the theory of connoisseurship or the apprentice model of guild knowledge; they create an environment whereby 'novice' judges can work beside and observe experienced examiners, albeit over a limited period of time, bringing their judgements to parallel the common standard.

Outlining the training meetings

The four modules from which participants for this study were drawn can be sorted thus:

	Voice Over Internet Protocol	*Traditional Physical Meeting*
Small (one team)	English 1	English 2
Large (many teams)	History 1	History 2

All meetings begin in the same way, with a standard briefing from an officer from Exam Board. This briefing instructs participants that the purpose of the day is for them all to reach a 'common well-founded understanding of the mark scheme as the Principal Examiner

directs', emphasising the hierarchical nature of the process, and setting the priorities of the day.

Traditional Physical Meetings

The normal procedure historically has been for all examiners to attend a one-day meeting, known as standardising or training, often in London, a few days after candidates have sat the examination, before they can begin marking. The Principal Examiner(s) will spend the interim time deciding on the application of, and setting the benchmark standard for, the generic mark-scheme in relation to the scripts which the exam has actually produced. For GCSE and GCE examinations it is not possible to pre-test the questions, so this is the first opportunity to see live scripts. There is also an opportunity at this time for schools to comment on, or complain about, questions and circumstances which they feel should have a bearing on the marking of the examination. In the case of larger meetings, a pre-standardising day is held the day before the standardising meeting, at which the Principal Examiner or Examiners pass on their judgement standards to the experienced markers who will act as Team Leaders during the marking period. This was true of both history meetings; both had face-to-face pre-standardisation meetings, despite the fact that one had an online training meeting. I was only able to attend the pre-standardisation for History 2. The key question at these meetings is 'can you justify the mark?' awarded to each standardising script, as it is Team Leaders who will have to answer the questions of the Examiners during the training process. For small cohorts of examiners the Principal Examiner(s) act as Team Leaders.

English 2

The smaller physical meeting was held in a conference room in an office building in London. The room, the size of an average secondary classroom, was arranged with groups of tables pushed together; all participants in the meeting sat around three double tables pushed into one large table. The meeting was opened by a Standardising Officer from Exam Board, who also registered participants and ensured that they had travel expenses forms and other paperwork. Once his briefing was delivered he left the room, although he remained available for support.

Actually taking part in the meeting were the Principal Examiner and his team of two examiners, one of whom is an experienced marker who acts as a Team Leader in the summer series. In addition the Chair of Examiners for the subject was present (who oversees

the Chief Examiners for each specification within that subject), and a Senior Examiner who had been involved in writing the specification, question paper and mark-scheme, but was not marking in this series. (She was also a Chief Examiner for another English syllabus). The participants all knew each other well and had a professional relationship stretching back some time. In such a situation the discussion ranges between all participants equally, and no-one is able to hide. Lunch and coffee breaks in the morning and afternoon all took place in the same room.

The Principal Examiner did not follow an authoritarian regime for this meeting. Discussion was egalitarian in nature, and there was no sense that the standard was unalterably set; the Principal Examiner was open to changing his views. Although he chaired the meeting, there was no sense that the examiners were being tested on their ability to reproduce his marking standard, but rather that they should come to a consensus on an agreed standard. This was the only meeting in which there was any sign of either a 'consensus' or 'contest' model of decision-making; the tone was for the most part, 'sweet reasonableness' (Christie & Forrest, 1981:35) but participants were unafraid to challenge each other's perceptions. The last word was had by the Principal Examiner who chaired the meeting; he took great account of other participants' opinions, however.

History 2

This meeting exemplified the traditional physical meeting. Held in a large hall in a conference building in London, it was set up with projection equipment at the front of the room, where the three Principal Examiners, the Chief Examiner and the Chair of Examiners sat at a desk off to one side when they weren't addressing the room. The rest of the room was filled with twelve or fifteen large round tables, each with eight to ten chairs. Each table hosted one team of examiners and their Team Leader. The paper had five different options available to candidates, and each team was assigned to one or other of the options, to limit the range of questions which they would be asked to mark.

At the back of the room tea and coffee was provided throughout the event. There were several representatives from Exam Board who registered participants, helped them find their teams, answered questions and distributed training packs. One representative delivered the Standardising Officer's briefing, which follows the same script for all standardising meetings from an awarding body.

The day followed a typical format: during the morning the Principal Examiners took the entire room through a set of common standardising scripts, which covered most of the

range of questions and attainment. Examiners read scripts to themselves, after which the principal examiner summed up the most striking features of each script, gave the mark it had been awarded, and justified the awarding of that mark. The entire session was carried out in plenary with no formal opportunity for examiners to ask questions, although their proximity to the Team Leaders enabled some queries to be answered instantly. Four entire scripts, each of two questions, were covered in this way. The Chief Examiner for the specification delivered a short briefing before the end of the plenary session.

Shortly before lunch control was handed over to the Team Leaders and from that point on each team ran independently and was able to move at its own speed and to leave the meeting when it had finished. Team Leaders took their teams through a further two scripts in the same way, both from the set of questions which the team would be marking. After these 'training' scripts the process moved on to 'standardising' scripts, where examiners were expected to read and judge each question for themselves, before feeding back to the rest of the team. The Team Leaders noted down the marks awarded by each examiner and asked for justification of the marks awarded before giving the mark which had been previously agreed by the Principal Examiner and explaining the reasoning behind it. Over the course of the day marking of the standardising scripts became quicker and examiners were able to approach more closely to the 'correct' marks previously awarded to scripts, at least those in the team under observation.

Voice Over Internet Protocol (VOIP) Meetings

VOIP meetings are arranged to echo the traditional face-to-face meetings. A 'registration' period begins the day, during which representatives from Exam Board check that all participants have the technology in place. Each participant has a microphone and speakers attached to their computer, and attends the training via a secure web-based log-in. Icons are available for individuals to indicate 'yes', 'no', 'laughing', 'applause', 'left the room' and 'hands up' without needing to speak; what they do say can only be heard if they hold down the control key on their keyboard. There was a slight delay because of the internet relay which sometimes made it difficult for participants to judge their turns correctly, or to understand what was being said because of overlapping speakers. The volume level of both microphone and speakers is controlled by the individual, which occasionally causes problems in communication or comfort.

The relevant documents are available for download as pdfs so that all examiners have access to the usual standardising pack, which they can either access before the

meeting from a website, or can download as required during the training session. The Principal Examiner has control of a virtual slideshow, which appears identically on all examiners' screens. The trainers demonstrated varying degrees of familiarity and ease with the online system, but most appeared to find it at least somewhat 'unsettling' (PE, E1).

English 1

A small team of approximately twelve examiners together with a Principal Examiner took part in this day long meeting. The session was conducted in plenary throughout because of the small number of examiners. Although it followed the same generic format of Exam Board briefing, then training scripts followed by standardising scripts, the distinction between the two types of scripts was blurred, so that little formal standardising took place. There was considerable discussion of the features of each script and the mark which it should be assigned, which was chaired and moderated by the Principal Examiner who was leading the meeting. She invited participants to talk, or gave the floor to individuals who showed an electronic 'hands up'. The marks which had previously been awarded by the Principal Examiner were, however, not negotiable.

There was a sense of fraughtness over the technology; it was the first time this paper had had an online standardising meeting. A considerable delay in beginning the training aspect of the meeting was caused by various team members, including the Principal Examiner, being unable to use the system or to access some pdf file. Once they had begun, the Principal Examiner was anxious to keep going without breaks so that further technological difficulties did not occur during a pause. The meeting ran smoothly once it had begun; the Principal Examiner announced her intention to run it as a traditional meeting and essentially fulfilled this ambition. The examination they were marking consisted of a few short answer questions followed by an essay, and is the only module featured in this study to include some non-essay questions.

History 1

This meeting was a larger one, composed of approximately 50 examiners and a number of Team Leaders, who were divided between the different historical periods which could be chosen as an option on this paper by schools. The technological difficulties went on proportionately longer for this meeting, given the larger number of examiners, so that the start of the training was delayed.

Schools (or 'centres') can choose from a number of different period options in this module, and consequently there were two Principal Examiners for this paper, who divided the different options between them, and thus the training. The first (PE1) is male and responsible for options A, B and D; PE2 is female and responsible for the other three options (C, E and F). They conducted an extended briefing on the 'lessons' from the previous summer's examining session, and then went through the training scripts, in the order of the options, swapping as necessary. A few questions were taken from the floor, but not many.

During a break for lunch the administrators from Exam Board assigned the examiners to the appropriate 'room' of the VOIP software, so that the afternoon session took place in individual teams, according to the option which they were marking, with one team assigned to each option for the most part. (One team marked two options which differed only very slightly in their content.) The Team Leader of the team I observed attempted to run the VOIP break-out session as he would a live session, which made for a somewhat stilted meeting given the constraints of the software and the difficulties which various examiners had with their equipment or connections to the programme. They covered the training scripts rapidly, with discussion severely limited by the technological difficulties that this team was suffering. The meeting concluded with the Team Leader's comment that he would telephone each of them individually over the next 48 hours to check that they were happy with the mark scheme, sample material and the procedures.

Training Strategies: Structural and Content Aspects

The review of the last session

The training from the Principal Examiners in all but one of the modules used the experience of the previous examination session to draw out lessons for the examiners, to suggest future behaviour, and to justify new rules or procedures. English 2 was the exception, since it was being examined for the first time in that session. In History 2 the review of the previous session did not form a discrete part of the training, instead being referenced in passing where it became relevant, so that the Chief Examiner on History 2 concludes the plenary training with a series of remarks prefaced with reference to the 'number of misconceptions that have arisen over the past year' before going on to correct them. In English 1 reference to the previous session was very limited, despite the Principal Examiner's assertion that as each question came up she would 'talk about how it went last time as I go and pick out pointers for marking and differentiation.' In fact the review of the previous session was confined to less than a minute's summary near the beginning of the meeting proper, which

characterised it as having gone '*extraordinarily* well', most significantly because examiners had 'got hold of the standards very very quickly' (PE, E1).

The two Principal Examiners for History 1 had clearly discussed the format and content of their training in advance, as they switched who was leading the session according to a pre-arranged order; they included a separate and lengthy review of the previous examination session for the benefit of their examiners. It was structured in two sections, beginning with consideration of the 'good points' from the summer before moving on to the 'bad points'. Some of the 'good' points are merely descriptive, for example that every one of the 72 questions was answered by at least some candidates, but the Principal Examiner then moves on the positive point that every examiner completed their allocation of scripts on time. The implicit message is that it is important that they should replicate this feat in the January session. This opportunity is also taken to praise and thank those examiners who contributed to the 'review process' (the re-marking of scripts after schools or candidates appeal against results).

The review then moves on to negative points, which dwell primarily on undesirable marking behaviours. This is interesting in light of the audience: examiners who are asked to return in January will not include those whose reliability fell below acceptable levels in the June examination series. This means that all those listening can apply the mark scheme within an acceptable margin of error, so the review of negative points is aimed at controlling certain behaviours in markers even if they don't lead to unacceptable levels of error within these individuals.

In particular the Principal Examiners for History 1 were keen to control the speed of marking, telling the examiners that 'it is much much better to mark at an even pace.' There is a discussion about 'normally utterly reliable' examiners who ended up having their marks for some scripts either dramatically increased or reduced; PE1 says that he can't understand this but 'it may well be speed or whatever.' There is a sense of warning them not to become overconfident or simply giving a reason for this control of the speed, so that examiners do not feel it to be an unnecessary burden. PE2 later reiterates this, asking the examiners to plan out their marking throughout the marking period so they can mark at a 'steady pace'. She also emphasises the idea that it is later in the process, when examiners speed up, that errors 'creep in'. This warning is supported by what might be over-strongly termed as a veiled threat that team leaders will be monitoring rates of marking through the reports on the system, concluding in the expectation of a steady pace of marking. This expectation of the Principal Examiners reinforces the system of milestones which are set and enforced by

Exam Board, so that examiners keep up with a rate which the board considers to be appropriate.

During this part of the meeting the senior examiners emphasise the need for reliability in that they want to make sure that examiners 'mark the very last script in the same way that we mark the very first one'. This is rationalised in terms of treating each script with 'the respect that they deserve'. 'Respect' is stressed and repeated by PE1.

Repetition is used to ensure the transmission of key facts on several occasions during this review; it may be unintentional, as these facts may simply be on the senior examiners' minds. Thus 'the scripts are no longer split' is reiterated several times in this portion of the H1 briefing, so that the review is also used to highlight what has changed in the process since the last time these examiners marked.

Introducing the mark scheme

For a session whose stated purpose is a 'common well-founded understanding of the mark scheme', it is surprising how little time is spent on considering the mark scheme in its own right, before scripts are considered. Examiners are expected to have familiarised themselves with the generic mark scheme before the training session, and some examiners at the live meetings spent the time before the training started reading the mark scheme and other training materials.

In only one module, History 2, was the mark scheme the subject of detailed scrutiny during the plenary. The Chief Examiner on that unit closed the plenary training session with a few remarks, which appeared to be generated by a fear of the 'misconceptions' which had been seen in the previous examination session, which caused her to spend some time going through the level descriptors for the different Assessment Objectives in some detail. Her main focus was on highlighting key words rather than defining terms; the difficulties associated with the mark scheme will be discussed in Chapter 6. Informal discussions among the team I was observing indicated that they had liked having a focus on the mark scheme and that it had been unusual, but that they would have liked it to come first in the meeting, before the Principal Examiner had given his commentary on the common scripts.

Since the mark scheme will receive detailed consideration in Chapter 6, it is enough to note here that there is no mandated structural focus on the mark scheme; other activities use the mark scheme, and it appears to be assumed that by undertaking the activities the examiners will unconsciously absorb the mark scheme. The intention may be to ensure their 'well-founded understanding' but in practice meetings seem to be more concerned with

ensuring the right decisions are made; if the system were a rational one as it is conceived in theory, the two would be the same thing, yet it could be read as a tacit admission that the mark scheme is not in fact the only tool being used by examiners.

Creating the framework

The training scripts provide examples of what aspects of the mark scheme mean, the standard which should be expected for a given grade and a range of exemplars for comparison. This creates a framework, both mental and physical, under which new scripts can be considered. With each subsequent decision a new point on the framework is created. The ordering of the plenary training can demonstrate an implicit awareness of this framework, and make it easier for examiners to take an overview of the material.

History 1 exemplified one strategy. The plenary training featured the following scripts:

	Level	Mark	Principal Examiner
Script 1	5	26	1
Script 2	4	20	1
Script 3	3	14	2
Script 4	2	8	1
Script 5	5	26	2
Script 6	5	30	2

These follow the options A-F on the paper through in alphabetical order, with the exception of the final two which are inverted. This suggests the ordering is deliberate, as does visiting the scripts in descending order before returning to the top of the range of marks. Selecting a number of scripts at the top of the potential range also suggests that there is an expectation: either this is the area where most examiners fail to position correctly – supported by the choice of two papers (1 and 5) which were awarded the first mark in level 5 – or that this is the area in which most scripts will fall. It is also possible that there is an intention to encourage examiners to award marks towards the top of the range.

History 2 followed a less clearly ordered format, possibly because there were fewer options on the exam paper. The training scripts were arranged in approximately ascending order; the first script demonstrated high performance in one of the Assessment Objectives, but was 'operating at level 1' in the dominant Assessment Objective for this module. The first Principal Examiner then gives a level 2 script, before his fellow Principal shows two

examples of level 4 scripts (the highest possible on this paper), the second of which is more secure than the first, and gains a few more marks.

The approach to the sample scripts, which examiners either marked in teams or individually before discussion, varied between the two papers. In History 1 there was no order demonstrated, but in History 2 the Team Leader introduced the break-out session by telling them that she had 'taken out the worst and best to look at together so that you've got your parameters', which also suggests the deliberate creation of a frame to delineate and guide the decisions made by markers. After this the 'classic horror trip', as she described it, had no pre-ordering to guide their decisions; such ranking would assist the examiners in making their judgements, but might be seen as giving clues which could not be relied on during actual marking. The training scripts seem ordered to provide a scaffold for the examiners to grasp, before they are required to take up the challenge of locating the position of scripts within the framework for themselves.

The scripts presented during English 2 (which had no division between training and sample scripts) showed no discernible order, although the Principal Examiner is concerned to check that 'we've covered enough of a range – twelve to thirty seven and a middle range', and moves across all of the questions, so that they have seen a sample of each of the questions from which a student can choose. There is a mixed approach in English 1; although the sample scripts for the first section on the paper are, as for History 1, bookended with top level answers, they display no order in between. For the second section, after the first answer the Principal Examiner comments that they 'don't expect these responses to be in strict descending order', apparently implying that one might normally expect that sort of a frame. Instead the first script is good, but not top level, followed by a completely top band script.

Ordering of essays has the interesting potential to create bias; it is conceivable that if the first response the examiner sees is very good, it would raise expectations for the rest, akin to the *representativeness* heuristic, and therefore leading to the under-rewarding of later scripts. There is clearly no established rule in operation between the units, and the organisation and deliberate thought of History 1 is unusual.

Training Strategies: Interpersonal Aspects

Although all examiners are employed directly by the Awarding Body, and receive uniform materials from the board, they really only form a personal relationship with their immediate supervisor; the Team Leader in the case of large examinations, or the Principal Examiner

direct in the case of small examinations. These individuals are also contracted by the board; it is not a full-time job for any of them (with the exception of some chief examiners and chairs of examiners, a level above principal examiners). The full-time employees of the exam board are administrative rather than subject or educational specialists, and have a very limited role in the training; the Awarding Body remains largely faceless and institutional in nature. As a result the individual's contact and relationship with their immediate supervisor becomes very important, in their performance management, in their carrying out of the process, and above all in the role the supervisor plays in training the examiners and creating their understanding of the process. One senior examiner told me during the lunch session that she found the face-to-face meeting essential to gaining a measure of knowledge of each person's character and trustworthiness. Reciprocally the team of examiners need to be able to put their trust in the team leader, to be happy at working under them and to be confident enough to ask questions, and thus support their learning. There are clearly challenges in regards to these interpersonal aspects involved in the VOIP training meetings, as so much of the communication is lost. While a senior examiner in a face-to-face meeting may identify which team members are failing to participate, and may therefore direct questioning and ensure the full involvement and participation of every member of the team, it becomes much harder to do so through the medium of VOIP. Still less may a team leader spot an expression of confusion or lack of understanding over the web interface.

It can be seen that the individual characteristics and teaching styles of the trainers become significant variables in the experience of examiners. While the overall structure and content of the training may be mandated at a higher level, Team Leaders have considerable scope for tailoring the delivery of that material. This section outlines some of the training strategies and approaches which were exemplified in the data. Some evidence will be offered to show that some informal strategies are as deliberate and officially sanctioned as the formal structure.

The two faces of the system

Two main approaches to training can be identified: the hierarchical, authoritarian vision and the friendly, egalitarian approach. The UK examining system is by definition the former, with standards set at the top of the chain and then passed down through cascading layers of training. Yet the general atmosphere of examination training meetings does not emphasise that hierarchy, despite the cautionary opening briefing. For the most part senior examiners rely on, and attempt to create, a sense of collegiality and an easy, open exchange of views.

The dichotomy can be easily illustrated within a single meeting: History 2. The Chief Examiner provides an opening warning before the Principal Examiners address the training scripts, telling the examiners that 'it is not fundamentally for you to question marks' and uses the imperative to stress that they should 'accept what God is doing!' The metaphor for the Principal Examiners is tinged with humour, but its serious intent was clear; her tone minimised the joke.

In contrast the Team Leader for History 2, while still giving firm instructions to her team about the submission of their sample of marking and contacting her, gives a much more friendly impression, reassuring her team for example that 'it is a difficult paper and very often people do a second sample – don't worry'. The use of the first person plural pronoun to refer to herself and the team as one unit ('then we need the six scripts that are [option] C scripts') suggests an egalitarian approach; this is a feature of the language of a number of other trainers, notably the Principal Examiner on English 2, and is considered further later in this chapter. (It is also seen in most of the meetings in the form 'and what do we think about this essay?', which is perhaps less egalitarian and more a remnant of the speaker's teaching days.) The Team Leader in History 2 also emphasises several times that a mark either side of what she (and by extension, the Principal Examiner) would give would not be a problem, so on one script she remarks 'I think you can move in level 2. If it's a mark or two in level 2, I'm not going to scream at you.' This does not reflect the authoritarian approach suggested by the beginning of the meetings.

This Team Leader also used a great deal of humour in her address to the examiners; for example at the start of the break-out session she tells the story of marking a paper which was filled with

> three or four pages saying that he hadn't revised the topic but that there was a really fit girl sitting next to him and he didn't want to look stupid so he was going to keep writing! (TL, H2)

This anecdote led to a general laughter and then the exchange of further 'war stories' which established a sense of collegiality, and functioned in another important way: the examiners might be called a team but few had worked together before, as they would be required to during this meeting. Mutual laughter demonstrates rapport and agreement (Adelswärd, 1989) and this suggests its function in creating a shortcut to mutual trust.

It would be easy to argue that the difference is that between a senior officer and a junior; the Chief Examiner, higher up the hierarchy, is more invested in the top-down

approach and will not need to interact with the examiners herself, whereas the Team Leader has not only to spend the entire day with her team, but also to supervise their work for the entire marking period, and liaise with them throughout it. However, in other papers the egalitarian approach is a feature of Principal Examiners too; notably the Principal Examiner for English 2. As noted above, this meeting is very small, and run entirely in plenary; it is also characterised by free-flowing dialogue in which all the examiners participate. Unlike in other meetings, there is no single voice which dominates the transcript. The marks which are given to scripts were all agreed between participants; the sense of a hierarchical communication of the standard was not in evidence at all. The Principal Examiner did not begin with a training script, so that the meeting went straight into reading a script and each person deciding a mark; he then asked 'can we have a discussion about what we think?' Here the first person plural seems to be a genuine one, not a disguised command, as he also enters fully into the discussion. The sense that he is in control is maintained; it is he who answers questions:

Examiner:	would it be too cruel to say it's limited understanding?
Principal Examiner:	I think so – I think that the word 'awareness' is key here.

Similarly the humorous tone of 'I'm sorry I'm picking out the difficult ones', while acknowledging the difficulty of the task does not suggest that he will make an adjustment as a result; this could be read as the use of a mutual challenge to unify the team, including the Principal Examiner, against a common 'enemy', a training strategy which is considered further below.

However, the flatter structure of the power relations in this unit is demonstrated by the fact that the Principal Examiner accepts the arguments of his team and adjusts the mark which he had previously awarded to one of the sample scripts; admitting 'I was thrown by the apparent brevity.' He can, of course, afford to make a change, where the Team Leaders on the larger papers cannot, as demonstrated by the Team Leader on History 1 who admits that the Senior Examiners were 'humming and hahing' over a script, but has to dismiss the arguments of his team in order to conform to the mark given by the Principal Examiner.

This admission of difficulty in deciding on a mark is a common feature of senior examiners' discourse, as a reassurance to the examiners, and with the effect of creating fellow feeling. The sympathy expressed by the Team Leader on History 2 that 'it's much easier when you're awake and away from all the noise' is a similar feature.

The rebel against the system

There are a number of places in the data where a senior examiner attempts to unify his or her team of examiners by placing them in opposition to a common challenge. The example of the challenging script during the English 2 training above is a potential example; similarly the Team Leader on History 2 positions them as approaching the difficulty of learning the mark scheme with hard work:

> Unless you try it... you can't do it didactically: you've got to [*makes motion with pen indicating the steps of a ladder in the air*]

In some places however, this effect is achieved by placing the system or the Awarding Body in opposition to their joint interests or abilities, inciting a minor rebellion against them, and creating a sense of conspiracy. In the case of the Voice Over Internet Protocol meetings there is a ready-made common enemy established in the form of the training format: the Principal Examiner for English 1, for example, tells her team 'I'm sorry that you can't talk to me, but it looks very difficult that you can talk [sic]', a sentiment which is echoed, though less explicitly by the Principal Examiners for History 1 also. The Team Leader for History 1 obviously has substantial difficulties with the system, commenting on it throughout the break-out session and concluding with an apology that 'it's been such a fraught experience.' He also suggests that he will telephone each team member individually to ensure they understand the mark scheme and standard – a personal touch which can have dual benefits in ensuring their understanding and making them feel valued. However, the VOIP system is only one example of something imposed by the Awarding Body on examiners, which they did not appreciate – there were mutters about the mandatory and prescribed introduction at the beginning of English 2, for example, and the quality of the venue and the food were both much discussed, particularly at break times, in the live meetings, with all levels of examiner participating in the discussion.

There was, however, one very striking example of this rebellion against the procedures of Exam Board. At the very beginning of the break-out session of the History 2 training, the Team Leader instructed her examiners to find the piece of paper titled 'Dates and Deadlines' – something which is a formal requirement of team leaders – and as they looked at the deadline for the first batch of scripts (interim deadlines for certain numbers of marks to be uploaded to the internet is a method of ensuring examiners mark at an appropriate rate) which was just two days away, she commented 'pigs just flew past' and went on to suggest that they should disregard this deadline. On the face of it this is a

rejection of the unrealistic expectations of the Awarding Body, and a serious lapse in her role as representative of the hierarchy of the examination system; such a rejection, coupled with humour, creates a strong bond with the team, and as I suggested, a sense of conspiracy, which makes the team more likely to accept the deadlines and restrictions which she does enforce, since she will clearly reject those which are too onerous for them. Yet this instruction, even the joke, came directly from the pre-standardising meeting which had been held with Principal Examiners and Team Leaders the day before. It was the Chair of Examiners who had pointed out the impossibility of meeting this deadline. Thus, apparently rebellious behaviour is in fact mandated, and at the same time provides an easy way to begin to forge a team bond.

The creation of a group identity

The use of humour, and the subversive rejection of some minor rule both contribute to a further strategy which is clearly demonstrated throughout the training meetings: namely the creation of a new, group identity. It is this group which has been seen as the 'community of practice' (Sanderson, 2001), but given the fleeting nature of the contact, I would argue it is more powerfully seen as a socialisation tool, which makes examiners happier in their role, and more willing to accept the supervision of the team leader.

Passing linguistic cues suggest the creation of this group identity. The use of phrases like 'you know' to introduce factual information create a bond between examiner and supervisor, flattering them as part of the group "in the know" as in the History 1 plenary. Similarly the Principal Examiners in English 1 when asking questions of the examiners do not use the second person pronoun but the first person plural, so that it is 'do we agree', rather than 'do you agree' as it is in History 1. Variation is present, clearly. It could be argued that it is even more important to use such linguistic cues during the online meetings, because the normal paralinguistic cues of body language and facial expression are not available as means to support interpersonal relationships. It is unsurprising that in the small team of English 2 examiners the first person plural is the most used pronoun, and instructions are couched as exhortations for everyone including the speaker: 'we have to mark discretely'. Even more blurring of the hierarchy is found in that unit (English 2), where the Principal Examiner is as likely to phrase instruction as if asking the advice of his team ('do you think with the "consistently analytical" we should be a bit lenient?') as to command directly.

The team leader in History 2 varies between 'we' and 'you', using the first person plural extensively, and confusingly, to refer to both the team and also the senior examiners.

There is one interesting exchange where she begins by telling them 'we're marking what they've done. We're not doing that thing where we diagnostically mark, that's not fair.' She continues to use 'we' and 'us' until her team embark on an extended argument amongst themselves about the accuracy of the 'debate' (which is the focus of the question), and she has to exert her authority, both to bring them back on topic and to stop them pursuing what she has already told them is bad marking behaviour: 'you can't negatively mark', she tells them firmly, setting herself outside their group.

This might seem like a tenuous argument, but it is important not to underestimate the extent to which the examiners are sensitive to such things. The mood of the room when, for example, the Chief Examiner of History 2, embarked on her lecture to them, exclusively using the second person and the phrase 'accept what God is doing', was markedly more hostile and less accepting. Examiners are adults who are used to being in charge, in their role as teachers, and granting them 'adult' status by the use of linguistic tags which associate them with those in authority is a way of minimising the friction caused by their subordinate role.

More explicitly, the Principal Examiners in the plenary of History 1 do something which is not seen elsewhere in the data. During the review of the last session, they praise examiners who contributed to the 'review process' (the remarking of scripts due to appeals from schools) by name, in the 'public' forum of the online plenary. It is the opposite of naming and shaming, and has more than a touch of the school assembly about it, but again there is the sense of creating a community, or at least making the community that exists less nebulous.

During the review of the last examination session one of the Principal Examiners of History 1 brings up the idea of respect for script or candidate (which is seen in every meeting), linking it to accurate marking. He argues that inaccurate marking 'wrongly persuades centres that the integrity of the examination is open to question, and that's unfair. It's unfair on all of us.' This suggests a concern both with maintaining the integrity of the examination but also on creating a common sense of 'us', as a community of experts or professionals, and the need to not let each other down. There is also a hint that he is attempting to draw on a sense of honour, and potentially its corollary - guilt, if they do not behave appropriately as professionals. The need to act as professionals is something which is stressed in the Exam Board standard briefing which opens every meeting, implicit in the 'code of conduct' which examiners are expected to follow.

To a certain extent, Awarding Bodies have an advantage in creating the identity of a group of professionals, as examiners are already teachers, with a strong sense of professionalism. The impetus for training meetings is to refocus that professionalism, and to establish the group identity as 'examiners' rather than teachers. Trainers do draw on teachers' established sympathies and attachment to students: the senior examiners leading the training specifically reference the people behind the papers - the phrase 'we have to ask ourselves, could a seventeen year old do any better than this' is familiar to any A level examiner, and it occurred more than once in these data. The need for accurate marking is put in the context of fairness to the candidates, and giving them what they 'deserve' so that one Chief Examiner commented as a decision was being made on a script that 'you're not doing injustice to the other candidates if you give this one 10' (History 2). That is, decisions on scripts are made with thought to the candidate who produced the script, and also to the others who sat the exam. The refocusing comes with the need to ensure that examiners do not overly sympathise with candidates, and the idea that accurate marking is equivalent to fairness. A group identity both enables the 'examiner' persona to override that of the 'teacher' and makes examiners feel more warmly towards the Awarding Bodies which are their temporary employers, and which can be all too impersonal.

Questioning

The use of questioning to teach, whether by guiding students to an answer or in a dialectical fashion, challenging arguments, has been called 'the most important feature of the classroom' (DfES, 2004:2). It is unsurprising that the senior examiners bring with them their teaching experience (and they are typically very experienced teachers) into the way they deliver the marking training. The questioning of examiner by senior examiners was an almost universal feature of the process, once the groups had broken out of plenary. In the main it was used for two purposes: for formative assessment, to judge how well the group of examiners were making decisions; and for teaching purposes, to explore the qualities of mark scheme and essays. In some cases the purposes melded, as examiners were asked to justify their decisions, which showed both what further training should focus on, and enabled discussion of the features which were relevant to decision-making.

Dialogic teaching methods were challenging to use during the Voice Over Internet Protocol meetings, because of the technical limitations. The importance of formative assessment had clearly been considered when the system was designed, however, as the symbols which attendees could display next to their names were heavily used to respond to

questions such as ‘is everyone happy with that mark’ as well as more mundane questions of ‘can you put a tick up when you have the B script open?’ The leader of the meeting also had the ability to control the page displayed within the meeting software, and the software was used to enable ‘voting’ for the placing of scripts in different levels, particularly in the plenary sections of History 1 and English 1. This gave the Principal Examiners a breakdown, in percentage terms, of how many examiners were in the right band, and how many were not, although it was an overall view rather than an understanding of where individual examiners were placing scripts. A similar overview was taken in the plenary session of History 2, during which the Principal Examiners asked for a show of hands on which level examiners would choose for the common script under consideration.

Despite the difficulty of discussion in the VOIP meeting, once the History 1 meeting had moved to separate team ‘rooms’, the Team Leader made a valiant effort at using questioning to guide his team through the training material. He used very open questions, typically ‘so what did you think?’, to stimulate discussion on the sample scripts; he directed his open questions at specific individuals, with the intention of ensuring that all examiners participated in the discussion, and would ask more than one examiner to comment on each script. It was unfortunate for him that most of the members of his team had problems with their internet connection or with their microphone or other equipment. One examiner participated via a text chat window, and he occasionally addressed questions to her that she could answer in that mode.

After an open question had been answered, the History 1 Team Leader often asked a follow up question. This was often just a request for a level, if he felt a sufficient commentary on the essay had been given. If a commentary was not sufficient in either length or clarity, then a follow up question might encourage the speaker to expand, with an aim to pin down the essential point for making the decision. The following exchange illustrates the typical questioning style of this senior examiner:

Team Leader: what did you feel about that one – what level would you have given?

Examiner 2: the m-main – the only problem was the lack of perhaps range that, the sort of depth, and the focus on the question means I find it hard not to give it a level five.

Team Leader: ahmm, now when you say range, what, you know, would you say was actually missing as far as that one was concerned? [*short pause*]

Examiner 2: umm enough discussion of other factors. It gives two in favour of the statement in great detail, and then obviously one, claiming Richard's rashness at Bosworth that doesn't really go into Henry Tudor's skills, well, though he sort of does. In the first paragraph he combines two points.

The Team Leader considers this with a lengthy pause, then asks another examiner for his view, which is essentially the same, before he then asks another follow up question:

> Did you think there was, either of you really, did you think there was weighing up going on about different factors in this essay?

Eventually, having failed to elicit the desired response from any of his examiners, the Team Leader sums up the official response to the script, which incorporates many of the things they have said, but also demonstrates the answer, or key word that he was clearly looking for, since it is the only difference between his explanation and that of his examiners:

> you know we felt that it just, needed more range as far as the stated factor was concerned, er to really make a stronger argument.

The key word required was the 'stated factor' while his examiners had been referencing 'other factors.' His extended attempts to get his team to provide the answer demonstrate some of the difficulties of the VOIP system; it may also have been that he was simply attempting to elicit an answer which was too far from any of their thoughts, and his questions are not leading enough to give the examiners clues to what they should be saying. Certainly none of the other trainers had so unsuccessful an attempt. The wait time after asking a question needed to be significantly longer in an online meeting, because of the mechanism required to speak, and the time lag of the internet. This was demonstrated successfully in English 1 but the Team Leader for History 1 rarely allowed enough time, and frequently answered his own question with 'Are you there?' and the name of the person to whom he had just asked the question, often interrupting them and causing confusion. This was, I suspect, partly to do with his discomfort at the system, and worry that his team members were not able to speak, as some did not have microphones, but it also highlighted

the difference communicative behaviours which are required for an online meeting. The Principal Examiner for English 1 used questioning to guide her examiners towards the correct mark; she would keep asking examiners for their opinions until the correct consensus had been established. Her greater level of success might have been due to her continuous flow of vocalisation ('hmm' or 'yup, yup') to direct responses, or the fact that she was simply looking for a mark from a very limited choice, rather than an unspecified historical factor from an unlimited one. For different reasons the Principal Examiner on English 2 also did not use questions as a training device, because of the more egalitarian structure of the meeting, and the free-flow of discussion.

The Team Leader on History 2, in contrast, asked for specific examiners' reasoning after each script had been marked and she had noted down the marks which had been given. Their responses were not usually extended and she did not engage with them directly, although there was some implicit evaluation. The exchanges followed the classic three part structure, of question, response and feedback, although the feedback was implicit rather than explicit. The Team Leader might rephrase something which had been given as an answer, but was more likely to simply give her own 'correct' answer to the question of what it had got and why. The following is a typical exchange. The team leader has asked Examiner 7 to justify the mark she gave a sample:

Examiner 7: it's well-integrated. Level 3. I didn't give it a level 4. I usually work down like this . Cos the reasoning, the way they use the sources it's not developed… it's integrated… it's weak but there's a discussion and the sources are still there from beginning to end.

Examiner 9: I think that the sources were used. There were nods towards using the evidence.

Team Leader: Sixteen plus eight. It's got a structure, it's got relevant own knowledge, you're not bowled over by the level. There's some understanding… it lacks depth.

They get a level 3 for own knowledge. Mary Seacole is fine, I like Mary Seacole. Florence Nightingale is beginning to pale on me! [*laughter*]

> It [The Sanitary Commission] gets mentioned but it really doesn't get analysed. One of the sources gets a kind of fleeting mention.
> Can't reward it both ways – we have given it 16 for integration in the first part.

The Team Leader does re-use words from Examiner 7's response, but does not prioritise them within her own explanation. She does not attempt to question her team further to elicit more of the summary of features which she is intending to give; she rarely uses follow-up questions of any kind. The impression from this team leader is that questioning is not being used as a teaching technique, but as an assessment, to diagnose how far her team are arriving at decisions by picking out the correct evidence. The next section will examine the way in which trainers give clues as to the correct answer during training, and it is perhaps not coincidental that this team leader is most striking in her use of this technique; potentially her cues are based on her assessment of her team's understanding, although it is not a link which it is possible to see in the text.

A trail of clues

Although examiners are expected to make decisions about essays from early on in the training meeting, usually beginning by finding the correct band on the mark scheme before progressing on to the finesse of specific marks, it is surprisingly rare for them to do so, even on the nominated 'standardising scripts' without being given either explicit or implicit direction from a senior examiner. The framework of scripts, described earlier in the chapter can also be framed in such as way as to provide some cues for the examiners. There are also a number of other times when what can only be described as 'clues' are laid by supervisors to enable examiners to make the right decision. I am not concerned here with the summation of the characteristics of the essay, the explicit direction or help which is given in the plenary sessions by Principal Examiners, which will be dealt with in the next chapter within the analysis of the running commentary which they provide on scripts. Such behaviour is not required in the kind of co-operative atmosphere of English 2, and on History 1 the team leader does not provide clues, instead using questioning as described above.

It is, however, a feature of the Team Leader on History 2 that she would give some kind of hint as to the nature of the script, as she did on all but the last essay which the team considered. For some she merely commented 'it's a nice script' or 'we're starting at the bottom', to give an indication of the level of performance. On another script she suggested:

> here I would say, look at the sources on question 1 and it's a really good example of why you must have the sources in the front of your mind. These are quite difficult sources... they're not easy sources (TL, H2).

It is easy to see that the 'key feature' for this examination was the sources provided for the student, but this extended comment suggests also that it would be easy to over mark this essay by assuming that the candidate had appropriately evaluated and understood the source material. Such a suspicion is confirmed by an examiner commenting at the end that 'on first reading it did look a bit as if it was better'; the Team Leader in her summation tells them that the essay received a low mark because 'they are completely confused with the sources.' By directing them to the most significant issue with the script, the Team Leader prepares them to succeed rather than fail – it is notable that she is keen to build up confidence rather than shatter it, and on one occasion rather than force her team to justify their marks when they have expressed a complete lack of confidence, she simply tells them the answer.

This team leader not only drops hints before her team read the essays, thus directing their reading and the things they are likely to attend to, she also on occasion gives clues immediately before the examiners have to commit themselves to a mark. On the second script she commented 'they're using the sources even if they're not doing the nice kind thing and saying "source 1 says"'; she then had to pause while everyone in the group took up pens and revised the mark which they had assigned. She laughed saying 'you've got the clue we liked it' (TL, H1). Similarly the Principal Examiner on English 1 tells her examiners that the scripts are not appearing in strict descending order she comments 'there's a little hint.'

The reasons behind this clue-laying seem to be two-fold. One is that it reinforces good behaviour, as the examiners learn to make the correct decisions and seek out the reasoning behind that, or to use the desired 'key features' as principal factors in their decision-making. A second could be linked to the sense of conspiring against the board, as discussed above in 'The rebel against the system'; the Principal Examiner for English 1 does half joke after her hint that 'I hope Sarah [Exam Board officer] wasn't listening!' The official process of the meeting is that examiners should make independent decisions, which they are assessed on, in that the marks they award are written down to be kept. Instead, the team leader mitigates insecurity by directing them to the correct level or mark, and creates a supportive atmosphere for learning, instead of a high-stakes test of their marking.

Feedback

Individual feedback to examiners is given after they have marked a 'live sample' of scripts from their allocation, an event which is outside the purview of this study. However, there is considerable feedback given to examiners from the outset of the training meeting, both in terms of the accuracy of their decisions and the features on which they should be concentrating to make those decisions.

The Voice Over Internet Protocol interface provides an opportunity for instant assessment and feedback that is unavailable in the real world, as when examiners have 'voted' for the band a script belongs in, the system automatically calculates the percentages choosing each band for the Principal Examiner. Only the History 1 meeting makes use of this feature; the Principal Examiners for that subject are much more comfortable with the features of the online system than the Principal Examiner for English 1, who makes frequent reference to the 'malarkey' of the system and how 'discomposing' she finds it. In History 1, then, the correct placement of scripts during the plenary session moves from 54% choosing the correct level for the first script to 100% choosing the correct level for the sixth, final script. The progression was not smooth, however, with the 93% correct on the third script diminishing to 73% correct on the fifth. The percentages were relayed to the markers as they went along, enabling them to evaluate not only how far they were from the correct answer, but how they were performing relative to their peers.

Although it is not a primary concern of the post-script commentary, there is some more considered feedback during this period. After asking for examiners to make a decision on the appropriate level, and making a comment on the ratio choosing each level, the Principal Examiner makes a comment on the rationale behind the correct level. The only feedback which is explicitly given is occasional positive feedback to the proportion of examiners choosing the correct level; on two occasions in the terms 'absolutely right' and 'correctly'. On one of those two occasions the second Principal Examiner's tone on seeing that one hundred per cent of examiners chose correctly also suggest strong approval. Once the second Principal Examiner also comments that it is 'lovely to see people thinking at, um, the top levels'. This praise also reinforces the possibility, suggested by the distribution of the common scripts over the mark range, that there is an encouragement to award marks towards the higher levels. Alternatively it is possible that this emphasis suggests there is a tendency to under-reward in general, and the training is attempting to circumvent this.

The Principal Examiner for English 1 gives similar feedback of 'brilliant', 'geniuses' and 'me too' when her examiners all get the right mark for a question. However, her use of

the system is slightly different, merely asking the examiners 'do we all agree this should have four out of four?' and counting the ticks which indicate agreement with this somewhat leading question. For the essay questions of the second section, she uses the ticks again, asking examiners essentially to put their hands up when she says the band they would place an essay in. Her wait time for the correct band is twice that for the incorrect band (*circa* 8 seconds as opposed to 4), although it might plausibly be that as people begin to tick, she must wait for them to stop. (The pressure to conform to the 'norm' of the group (Bettenhausen & Murnighan, 1985) could also be in operation here, as examiners can see the growing proportions of their colleagues who have chosen the level.)

It is easy to see that such a system cannot be in operation during the plenary of a large meeting, although the History 2 Principal Examiners frequently ask for hands up to indicate which level examiners would award a script, and this provides some of the same kind of instant feedback, with the added aspect that the team leader, sitting at the table, can see and assess how her team have judged the essay. In the English 2 meeting it would have been against the ethos of the meeting.

Although it might be expected that the break-out meetings would provide more scope for feedback more tailored to the examiners' particular comments or reasoning, there is little evidence of this. Comments from the trainers tend to be addressed to the entire team, rather than specific examiners; sometimes these comments are prompted by something that a single examiner has said, in the manner of feedback, but made into a lesson for the entire team. Thus in History 2:

Examiner 2: I wrote the correct comment but still gave it a higher mark

Team Leader: go back to what they're supposed to be doing – look at [the mark scheme]

The Team Leader's comment was addressed to the whole team, rather than just Examiner 2, and the entire team at this point went back to the mark scheme. Similarly in History 1 the Team Leader made comments which implied positive feedback, by incorporating elements of the summary provided by his examiners, acknowledging these elements with the phrase 'as one of you said.' Specifically tailored feedback of the sort that suggests individual examiners are being too harsh or too generous is reserved for the more formal marking sample procedures. It is possible that giving such specific feedback during the group meeting would be counter to the positive atmosphere which most trainers seem to try to create; it is also possible that it is simply too difficult to keep track of individual examiners' scoring and

supply feedback during the meeting. Most examiners will give themselves some sort of feedback during the meeting and then use it to correct their marking after, as they note that they have been much more generous earlier, for example, and automatically err on the side of severity for the latter half of the meeting; this was demonstrated by a number of different comments spread across the History 1, History 2 and English 1 recordings. It would have been out of place in English 2.

Conclusions

The character of the training meetings and the strategies used to inculcate required behaviours, or Sanderson's 'cultural heuristic' (2001), vary dramatically between different trainers. The Principal Examiner (or sometimes the Chief Examiner if present) sets the overall tone for the meeting, and it is noticeable that the larger meetings are more rigid in their adherence to the marking hierarchy, as claimed by Sanderson. Team Leaders also vary in their behaviours, and the contrast between the two team leaders for history demonstrates that different strategies are employed, possibly unthinkingly, by different trainers even within the same subject. The use of questioning and of feedback, two key pedagogical techniques, have been noted within the standardising meeting. In addition there are a number of features of the language of trainers which suggest an attempt to create good interpersonal relationships between examiners, and to create a sense of professionalism which is linked to accurate marking. Online and face-to-face meetings have different challenges and require different linguistic approaches, particularly in the matter of turn-taking and dialogic learning; the VOIP interface is one which unsettles many trainers and examiners, and makes their relation to one another problematic.

One of the key structural aspects of the training meeting, the commentary on essays by trainers, has not been explored in this chapter; it is considered in the next chapter, along with other evidence of detailed analysis of scripts, and the factors within them that engage examiners' attention.

Chapter 5 - Analysing the scripts

"When the mind is thinking, it is talking to itself." Plato

If it is necessary to ask examiners to 'think aloud' to gather data on their decision-making processes, during the training meetings they do it without being asked. A large part of the training meeting is devoted to providing a model for the critical analysis of scripts which forms the basis of the application of the mark scheme. The running commentary provided by senior examiners is almost 'think aloud' in nature: it highlights the thoughts which present themselves to a senior examiner on reading, though edited to be of the greatest benefit to the other examiners, as their supervisors see it. Thus the senior examiners model their cognition about, and response to, any given essay.

The commentary on scripts demonstrates the features which draw the attention of examiners while they are making decisions, therefore answering one of the sub-questions driving this research; this chapter describes the foci of attention that can be identified. It also considers the comments which senior examiners make on scripts and categorises them. A range of types can be identified, from the straightforwardly descriptive, through judgemental, both implicit and explicit, to the extrapolatory, which take specific markers and apply them to the wider context. A typology of these comments is suggested at the end of this opening section.

Analysis of scripts is a prerequisite of applying the mark scheme. To decide if a script matches a criterion such as 'uses a wide range of relevant terminology', for example, the examiner must establish the range which is used, how relevant it is and how wide it is. In that respect, during the analysis of scripts examiners seek to establish both what is done by candidates, and how well they are doing it.

Three of the meetings followed the same basic format of script analysis: the plenary session at the beginning of the meeting involved the Principal Examiner or Examiners analysing aloud a series of scripts, with or without formal judgement being required of the examiners; then in the break-out sections Team Leaders typically asked their teams to undertake analysis, before or after giving an approximate mark, and then gave the agreed mark and a summary analysis which explained the judgement which had been made by the senior examiners.

The anomaly is in the meeting for English 2, in which there is no model analysis as such provided by the Principal Examiner. As discussed in Chapter 4 above, this training

meeting was run as a democracy rather than a dictatorship, and resembled the break-out sections of the other meetings much more than the plenary sessions. That is to say, a group analysis was carried out, on the instigation of the Principal Examiner, and the other four examiners participated on a reasonably equitable basis in the discussion. He, however, did not contribute much to the discussion of the scripts, withholding judgement, asking questions which did not appear to be leading, unlike those in the break-out sections of the other papers, and in fact appearing almost hesitant in his own judgement. The presence of two more senior examiners in the meeting may well have contributed to that. When he did contribute analytical discussion it was precise and referenced.

As with many of the features of the training meeting, the ways in which analysis is carried out or modelled depends both on the personality and preferred style of the individual Principal Examiner or Team Leader, and on other contextual factors, such as the size of the meeting. English 1, for example, was a Voice Over Internet Protocol (VOIP) meeting, but the Principal Examiner nevertheless made attempts to open up the floor to examiners to analyse the scripts, which was possible because of the small number attending the meeting, although the constraints of the system, and the participants' lack of facility with it, made discussion difficult.

What is analysis and what is merely commentary?

The level of analysis demonstrated by examiners pleasingly reflects the level of analysis in the candidates' own answers; in both English and History essays aim for analysis but sometimes fall short, operating merely at the level of description. Similarly, examiners may intend to analyse, but often simply state their own reactions, which are sometimes purely emotional, or make comment on the sentence or paragraph which has passed before them. Analysis requires separation of a whole into elements, the identification of those elements (Chambers, 1998) and then, in the case of examination, a value judgement. It is hard to distinguish whether a case is analysis or commentary, until the two are juxtaposed, as they are in this study when the different training meetings are considered together. Ironies often arise in the transcripts, as examiners criticise scripts for not using enough supporting detail, while they themselves make general comments without substantiation from the script under consideration.

The distinction between analysis and commentary can be explored further. Analysis of the data demonstrates that a number of different types of comments were utilised by the

senior examiners (principals and others). A potential system of classification drawn from the data is suggested here, and further explanation of each type given immediately below:

1. Descriptive; non-judgemental
2. Straightforward implied judgement
3. Explicit judgement
4. Comment containing analytical extrapolation
5. Identification of signals
    a. False
    b. True
    c. Uncertain

The first type is non-judgemental, in simply noting relevant elements in a descriptive way, such as 'so we've got the given factor addressed immediately'. The 'immediately' is not specifically a hint to judgement, but the fact that the word is used suggests that it is significant. Within this type also fall comments which involve low level analysis of the structure of the essay to enable the conceptualisation of an overall shape to the essay: 'then a concluding point'. This meta-signposting of the structure of the answers does not usually mirror the signposting of the discourse cues used by the candidates; on occasion, however, the commentary by the first Principal Examiner in History 1, for example, echoes it, usually in advance, so that we get passages like:

> then the candidate turns to other factors*: «other factors were advantageous. Other factors helped William to win the battle »* and we turn to Harold Hardraada and the invasion of the North which is of the double invasion from North and South*. «If it wasn't for Hardraada and his double invasion then Harold's forces would not have been depleted. »*

To a certain extent the repetition is redundant, and the Principal Examiner seems to be aware of this as shortly after he reduces the tautology by only giving his descriptive summation of what is going on, and leaving out parts of the essay that prompt it, so that the candidate's sentence *«these were the main factors which influenced William's success, though others to think about are luck and the view of the battle as a Holy War»* is presented orally as 'the candidate says these are the main factors, and then says look there are two others - luck and the view of the battle as a holy war' (PE1, H1). These type of comments are common when the senior examiner finishes reading a script, as they summarise the shape of

the essay. It is surprising how rare descriptive comments with no implied judgement are; there tends to be a judgemental tone even in noting, for example, that a script moves from one text to the other ('it's someone who's been told you have to' (Examiner* in English 2).

This second type contains judgement in the tone and expression of the comment which accompanies an extract from the script: *'« William's fleet and troops left in September rather than earlier. This was a good military tactic.»* Well of course it wasn't, it was just good fortune' (PE1, H1). In this case a judgement is suggested by the direct contradiction of the candidate's statement. In other cases the vocabulary used demonstrates implied judgement: 'the candidate then comes to a conclusion which attempts to evaluate what has just been said' (PE1, H1). The use of the word 'attempts' implies that the candidate was unsuccessful.

In other places the judgement is not implicit, but stated straight out (type 3): 'basically the comments show that they're not very secure' (PE, E1). In other places specific examples are picked from the script and judgement made: 'integration: it's quite a sophisticated historian thing to do' (TL,H2). The way the essay is constructed is labelled ('integration') before the judgemental comment.

The judgemental comment has much in common with the fourth category of commentary: however, this type goes further, from the specific to the general. In this type the senior examiner models analytical consideration of the candidate's script, so that an extract referring to William's *'«well-organised fleet »'* is reshaped and produced for the examiners as consideration of 'logistics as a military quality'. The Principal Examiner used an example from the script to place it into an overview of factors, drawn out to the level of category rather than simple specific incident, making it a more generalisable feature and thus more useful to the examiner.

Most of the comments which fall under the final type refer to various factors which might or might not function as heuristics, which are discussed below in Chapter 8. This identification of signals takes fragments of scripts, usually in terms of specific phrases, and links them to the overall quality of the essay, also in some ways an extrapolation. The signals tend to be mentioned when they suggest the quality of the essay to be other than it is, although some true signals are also identified by senior examiners; at other times signals are identified but judgement is reserved as to whether they are true or false. The point of identifying such phrases is clearly not only because they are of relevance to the specific essay but also more generally, which may be why the uncertainty is preserved. Such signals include references to discourse markers, plans and Quality of Written Communication.

This typology is relevant to the discussion in the main part of this chapter.

The discussion of the analysis of scripts at training meetings below has been divided into two parts, the plenary and the break-out sections. These two sections, as discussed in Chapter 4 above, have different purposes and the distinction has provided the most convenient way to break up the data. The alternative is to consider separately the analysis provided by Principal Examiners, Team Leaders and examiners, but the lines of discussion are rarely so clear-cut. History 2 is unusual in following what might be thought of as the norm, which is a plenary session exclusively featuring the voice of the Principal Examiner. In contrast, although English 2 is entirely composed of plenary, the Principal Examiner guided analysis rather than providing it, and the characteristics of their dialogue much more closely reflects those of the break-out sessions.

Plenary

Typically plenary analysis includes the reading aloud of the script by the Principal Examiner (in all modules except English 2), which enables line-by-line, or at least section-by-section commentary, which in some cases reaches the level of analysis.

In History 1, the first Principal Examiner used his commentary to create an outline of the answer, providing the kind of simple discourse markers which are not used by the candidate: 'so we've got a second factor that's brought into play and assessed.' He moved between anaphoric and cataphoric reference; comments such as 'so you think ah good there's going to be other factors highlighted here but they're not' direct the response of the examiners to future content, and demonstrate a response to individual lines in the context of the answer as a whole. This Principal Examiner created a conceptual map of the answer by his commentary, which enabled easier application of the mark scheme at the end of reading. He therefore fulfils the criteria of 'model' in at least one way, in demonstrating an approach which is advantageous to examiners.

The second Principal Examiner in History 1 also did this to a lesser extent, signposting the content of a script ('[it] then moves on to individuals'), but her progressive analysis was more concerned both with what the candidates were doing and how well they were doing it, a distinction which will be discussed further below. Both Principal Examiners noted inaccuracies or incorrectness in passing.

Each of the two completed their analysis with a brief summation, which sometimes summed up the characteristics of the structure of the essay ('we have here an answer which

makes five points, has secure quality of communication and has a conclusion' PE1) and sometimes merely restated the content which had been covered. After this summation the examiners were asked to place the script within a level, before a more thorough summative analysis was given to explain why the script was awarded the mark it actually received. The second Principal Examiner tended to give a more balanced and thorough pre-levelling summation than the first; she might end her reading of an essay with five or six summative statements, some which are positive, some which are negative, some of which reflect precise analysis, so, for example, '[it] does not address the economic difficulties even though it says in the beginning that it agrees to a lesser extent'. The first Principal Examiner in contrast was more thorough in the post-levelling summation, and described the characteristics in terms which are more closely related to those of the mark scheme: 'if you check that against the level 4 descriptor - analyses and shows some understanding, supported by some accurate factual material mostly relevant'. Principal Examiner 2, on the other hand, did not tend to use the language of the mark scheme even post-levelling, although since she was still considering the skills demanded of candidates, she might refer to the 'very basic understanding' demonstrated by one essay.

In contrast to the Principal Examiners in History 1, the Principal Examiner for English 1 did not usually operate at the level of analysis in her line-by-line reading and discussion of the scripts. The vast majority of her comments were immediate responses to specific statements, expressing value judgements, either in words ('that's actually quite perceptive') or frequently with non-verbal cues, 'hmmm' or a laugh. However, this apparent tendency to commentary rather than analysis may be a result of the fact that she interpolates far more comments into her reading of the script, saying something after every couple of sentences, at least once a minute, as opposed to the Principal Examiners in History 1, who will usually comment only once every two or three minutes, after a substantial paragraph of content.

There is also analysis in English 1: a few times per script the Principal Examiner will make a response which characterises what the candidate is doing at a particular point in the script: 'so it's embedded the analysis there, in quite a fluid way.' These sometimes appear to be making points which might be overlooked or dismissed by examiners: 'a bit wordy but it's talking about subject specific lexis.'

In addition in English 1 a more summative form of analysis appears, which draws conclusions about the entire script, expressed occasionally through the commentary, often using the terms of the mark scheme: 'that's competently structured and quite insightful – hitting loads and loads of AOs'. The Principal Examiner did not summarise the overall

characteristics at the end of the essay, preferring to open the floor to discussion, which she could do because of the small number of examiners; instead she stated her conclusions at the point of the evidence in the text. This is not model behaviour, in that the Principal Examiner might know the shape of the overall text, and could therefore draw conclusions mid-essay, but her examiners would not have known if a characteristic was typical or extraordinary until they had read the entire script, and were usually, therefore, discouraged from drawing conclusions too quickly. The discouragement, seen in several modules, might take the form of warnings that scripts 'look' as if they are going to do one thing but do another, or implied multiple earlier readings by Principal Examiners.

The Principal Examiner of English 1 seemed to be less systematic in her approach to the analysis of scripts than her colleagues on History 1; it is tempting to speculate that she could afford to be, because of the nature of her audience, who are entirely accustomed to the analysis of texts for key features, by virtue of their subject. (Although historians are also analytical, and consider themselves to be so, mentioning 'analysis' many times as a key requirement of candidates.) Certainly the examiners on English 2 would support the theory that English examiners are able to analyse features of scripts competently without the need for a model: the Principal Examiner on that paper offers little guidance, yet all four participate happily in a group analysis of the sample scripts. The single difference between their analysis and his is that while they make summative, overall judgements (e.g. 'it's fluent'), he tends to anchor his analysis with specific evidence from the script, directing his examiners to the 'sophisticated' 'engagement with writer's literary and linguistic devices' demonstrated by a candidate's discussion of the word 'gossamer'.

All Principal Examiners identified in passing, usually during the line-by-line reading, errors in the candidates' answers, either with an explicit 'that's wrong', or other comment, or a non-verbal 'hmm' or laugh. They were not usually further explored. The errors which attracted attention included factual mistakes about, for example, dates, misspellings and more complex issues, such as a complicated false assumption about the characteristics of *Guardian* readers in English 2 (which did occasion greater discussion). This demonstrates that errors do catch the attention of examiners as they read through, creating a momentary hiatus in the smooth processing of a script's contents.

There is no line-by-line analysis by either Principal or other examiners in English 2, as they read the scripts to themselves. Similarly during the plenary section of History 2, the examiners were required to read the script to themselves, before the Principal Examiners summarised its characteristics for them and linked those characteristics to the mark scheme.

The characteristics which they touched on in that meeting ranged from the style of writing, the content in terms of context, attribution of sources, the relation of answer to question, and the approach which the candidate had taken, usually expressed in terms of structure. The second Principal Examiner on History 2 also supported his analysis with specific reference to the script, very necessary when there has been no shared reading, and creates more of a sense of a line-by-line analysis by moving examiners chronologically through the salient features of the script: 'it does, if we go on to page 8, attempt a more explicit analysis of the source.'

All the principal examiners demonstrated a focus, as would be expected, on the key factors for the paper, as required by the Assessment Objectives. Thus the 'stated factor' and 'other factors' in History 1, the sources in History 2, and 'context' in English 2 appear regularly in the analysis. This is most obvious in the History papers, which have very specific assessment foci, while both English papers require literary and linguistic analysis of the text whatever else is a focus for the essay. In History 1, for example, the analysis of the first common script mentioned the word 'factor' no less than twelve times in under eight minutes, without including the number of times it appeared in the candidate's essay. This repetition reinforces the centrality of the key factor, and it was explicitly noted as a hurdle which must be cleared to get higher than 18 marks, a comment considered below in Chapter 8.

Not all scripts received the same degree of commentary during the plenary session. Some scripts were left to speak virtually for themselves, with far fewer interpolated comments and with much shorter summations – so that the first Principal Examiner on History 1, on his last script, can follow the statement 'that's the end of the answer' immediately with 'what would we give it?' It was considered to be obvious, and in this case it was because the script is so poor; he appears justified in his judgement, since 82% of the examiners, according to the computerised monitoring system, correctly identified the level of the essay. Similarly the second Principal Examiner made far fewer comments on her final two papers, aside from directing the audience response with the words 'you should get a sense of the level almost immediately'; these two were both very good scripts, securely in the top half of the top level. Very good scripts require less outlining to allow the reader to follow them, so comments that do occur fall into the category of judgement with implied analysis, so we get 'good or excellent chronological understanding' (Principal Examiner 2, History 1). In addition very good scripts tend to feature strong signposting, so that the meta-commentary outlining

of the answer which is sometimes seen is not required. Thus both very bad and very good answers are seen to need less thorough exemplary analysis, since their features are more clearly obvious to the reader and require less excavation. Placement in the middle bands, or just into the top band, is a more challenging task, and more careful analysis of the scripts, and of the gradations of their achievement, or at least analysis at greater length, seemed to be required.

Using analysis

Principal Examiners use the analysis of the common scripts as a frame on which to hang other key learning points related to the marking of scripts. The second Principal Examiner on History 1, for example, identified a discourse marker that suggested there would be discussion of the 'given factor' before commenting that she does so 'bearing in mind that an answer that does not address the given factor cannot go above 18 marks.' This rule, unusually hard and fast in its clarity, was then referenced several times during the break-out meeting by the examiners.

Similarly the second Principal Examiner for History 2, aware that his analysis had used the word 'provenance' several times, and that he had picked out and praised its use by candidates, told the examiners in an aside during his analysis of a paper: 'you may be getting drawn into the idea that you always need to use the provenance to weigh up – and you don't.' This reflects another characteristic of the analysis - the warning. Principal Examiners in both History modules and English 1 all identified what might be described as 'false friends', either in the characteristics which they were picking out, such as this example of provenance, or in the signals picked out from the text. (The English 2 meeting also discussed them.) This extrapolation from the specific example to the general application to marking makes explicit the reason for analysis. That is not to say that other comments have no relevance to the wider picture: examiners can and do extrapolate themselves, particularly in the use of comparison as discussed in Chapter 7. However, the extrapolatory comments by senior examiners emphasised key lessons or make specific warnings to prevent the formation of false conclusions about general rules: they intervene at significant points, rather than at all points.

To this end all the Principals identified signals in the writing of candidates which should attract the attention of the examiner. The Chair of Examiners during the plenary session of History 2 commented that it was 'important that some things raise a banner in your brain.' In English it might be the fact that 'some of the quotations are nicely embedded'

(Senior Examiner, English 2); in History it might be the elements of the introduction that indicated a 'multi-factored approach' (Principal Examiner 1, History 1) or the fact that an answer 'uses the sources as a set is a signal this is going to operate at a higher level' (Principal Examiner 1, History 2). The recognition of these signals is an element of the analysis which demonstrates that it has a wider function than merely analysing the script at hand for the purpose of assigning it an appropriate mark. There is also a level of analysis above that, of what features are reliable, and which are not, of what has a broader application and what is limited to the essay under consideration.

One of the advantages of using signals is that they form a shortcut for the comprehension of the answer, addressing the issue of memory capacity, and bringing an element of simplicity to a deeply complex task. But it may be too tempting to rely on a simple signal, a key word or sentence that indicates the presence of something which might not actually be present. Allowing, for example, discourse markers, to become proxies for actual features is unwise; as a result the reliability of these signals, and the risks of relying too heavily upon them was also a theme in the commentary of Principal Examiners during their script analysis. The comment from History 2 quoted above, about the use of the sources as a set was immediately followed by the caveat that 'because they know they should, they say they're going to, but don't' (PE1, H2). The word 'deceptive' or variants thereon, arose in almost every meeting; without perceptive analysis examiners might fall into the deception and take scripts at their own face value. This hedging of the suggestions made during the analysis relates to the unofficial heuristics and their application which will be considered in Chapter 8; the deliberate creation of uncertainty will be considered further in Chapter 10.

One particular 'false friend' which is identified with some regularity is the issue of 'Quality of Written Communication'. Nearly every Principal Examiner referred to it regularly, either in so many words or as 'language', in passing through the essays, in line-by-line analysis, in post-reading summation or in the discussion afterwards. These very frequent mentions belie the warnings against relying on language as a marker of the quality of the answer, and the insistence that Quality of Written Communication is a 'marginal matter' (PE1, H1).

The Principal Examiners used the plenary modelling of script analysis expertly to indicate which features they considered to be important in the judgement of scripts, and how those features interact together. The script analysis during the plenary is designed to demonstrate

how answers may be deconstructed in order to match their characteristics to those described in the mark scheme, and by reducing the long answers to smaller pieces, in sentences, paragraphs or ideas, the exemplification of the contents of the mark scheme becomes easier. The characteristics are established in relation to the question, in relation to the wider context of candidates' answers, and in relation to the mark schemes.

Break-out

The script analysis during the break-out sections has another purpose from that of the plenary: to allow the normal examiners to attempt the analysis they have seen modelled with the guidance of a senior examiner, with commentary forming an abbreviated analysis supplied after each exercise by the Team Leader, to allow them to adjust their own behaviour. The scale of the break-out sections is much more manageable for the larger meetings, allowing more questioning from both the trainers and those they are training, affecting the responses which are given by both sets of participants.

The balance between analysis by Team Leaders and by examiners is different for each module, just as all the training strategies vary from person to person. Three of the modules (History 1 and 2, English 2) have examiners reading the scripts to themselves (where appropriate English 2 will be discussed in this section, despite the fact that there was only one team of examiners). In History 1 the Team Leader asked a team member for their analysis of a script, which he then rephrased as necessary, and then asked a second member for their comments. Finally the Team Leader gave his analysis. For the first half of the afternoon he gave his analysis, then asked them to give a level to the script, before finally giving them the mark it received, and any justification required. Mid-way he switched to asking them for a level before giving his model analysis, which was accompanied by a mark and an explanation of how the senior examiners arrived at that mark.

The Team Leader in History 2 gave commentary on the scripts while her examiners were reading them, guiding their expectations. She then asked for a mark from each examiner and on the basis of that mark chose examiners to interrogate for their analysis of the script. Further discussion took place around the edges of this format; the live meeting provides for far more easy discussion and argument. After the examiners had attempted to justify the mark they gave, the Team Leader gave the correct mark and explained why that mark was appropriate, providing in some ways a model for the annotation rather than for the analysis of the script.

During the break-out sessions a much greater range of foci of attention are identifiable; some of these foci were suggested by the Team Leaders, explicitly or implicitly, as being worthy of attention; others came from the examiners themselves. This difference suggests the way in which the Principal Examiners led their audiences towards the features which they considered to be good prompts to judgement, rather than allowing a wider range to confuse or distract. The examiners, however, wanted to question the Team Leaders on the appropriateness of noting a particular feature, or how far another mattered. In History 2, therefore, the following exchange occurs:

Examiner 7: «*Adds weight to the argument»* it says.

Team Leader: But that's not really analysis.

Examiner 10: What would be analysis?

Team Leader: It may be that they're analysing up here [*indicates lines*]. It may be that they've got their argument and they're just grabbing. They wouldn't be the first historian to do that.

Specific examples from the scripts are an important part of the process of analysing the script, and this Team Leader is particularly expert in it. The question of what constitutes analysis is a particularly important one for examiners in both English and History. One of the key distinctions in the History modules in particular is that between an analytical answer and a merely descriptive one; the question also arises in English 2 with great frequency. Analysis is seen as a higher level skill which is essential for both English student and historian: the Team Leader on History 2 comments at one point that a script 'does slip into comprehension – it's not a history paper.' The implication might not be one which an English teacher would welcome, but it does demonstrate the skills which are valued. At one point the Team Leader on History 1 gave an examiner clear confirmation, a rare event, that analysis is 'a turning point', and that any sort of attempt at analysis gets the script beyond a certain mark. An examiner on History 1 made the distinction between analysis and description clear as he comments '[it] attempted analysis – quite a lot of descriptive passages.'

One of the major themes which can be identified in the examiners' analyses is that of Quality of Written Communication. The quality of writing is frequently mentioned, and taken to be an indication of the quality of the answer in general, despite the insistence by senior examiners that it is a 'marginal matter' (Principal Examiner 1, History 1). One examiner took the opportunity to ask during a plenary what effect the Quality of Written

Communication should have on a mark (History 1), and another asked the Team Leader during the break-out session in History 2 how they should 'penalise' for poor spelling. The Team Leader replies that 'if you're dithering and it's appallingly spelt go a little down, but it shouldn't shift it a level.' Despite this almost every examiner at every level made some sort of comment about writing style, fluidity, or expression. The Team Leader during the English 1 break-out session commented in passing through a script, for example, 'that's true, it just lacks the fluency to communicate that to us clearly.' Others mention 'one or two inaccuracies [of spelling and grammar]' (PE, E2) or the misspelling of the word Quakers as 'quackers' (PE2, H1). In English it is not always easy to establish what is a relevant subject domain-specific skill and what should be characterised as Quality of Written Communication; the word 'vocabulary' for example might be seen to refer to the range of words used, and therefore be a writing issue, or to the technical terminology relating to literary and linguistic devices, which is a subject skill.

In general, however, the main focus of the analysis remains on the key factors of each module, so that the 'stated factor' or the context or the literary and linguistic devices form the backbone of the analysis for all examiners, not just those leading the training.

A balanced view

A major difference in the analysis seen in the break-out sessions rather than in the plenary is that examiners tend to be more balanced in their analysis, in that they relate both the bad and good features of a script. An examiner on History 1 summarised an essay in the following way:

> the main problem was the lack of military discussion, but the point about Margaret of Anjou was quite good, in terms of the mistakes and how she brought things on to herself. And they never explicitly say "this is a political mistake" which you probably wanted them to, to make a more focused answer but, as you said, it's sustained analytical approach all the way through. The material's not really fully under control, but it is a good attempt (E3, H1).

Each bad point or 'problem' is counterbalanced by a feature which was good, a balance which is rarely seen in the senior examiners' summations. This may be because Principal Examiners know the end to which they are aiming, and lead their audience through the script by picking out the features which direct the reader to the judgement which the Principal Examiner wants them to make, whereas the examiners are trying to find their way

blindfold. The balancing between good and bad suggests a weighing up of the two sides, in an attempt to find where the pivot – the mark – should be. However, the analysis of the script is not only a tool to help the examiner come to a judgement, it is also in some ways a test of their performance, as they must tell the Team Leader what they think of the script and show how well they have absorbed their training. A balanced approach also hedges their opinion, meaning that they can take further cues before coming down one side or another.

Balance is the key characteristic of the entirety of the analysis of English 2, not in any one individual's extended turn, but in the discussion by the entire group. A short extract from the discussion of one script serves to illustrate this point:

Examiner:	It's evaluated - [we] haven't seen that yet.
Chief Examiner:	It's consciously structured to support an argument.
Examiner:	You could quibble that perhaps it doesn't have the detail. When it does so it's very good.
Principal Examiner:	I thought the way the contextual detail is developed at the beginning and developed into a quotation… on page 5 they start discussing the context and losing sight of the question

The participants analyse the good and bad features with rapid changes of turn, while each person is capable both of praising and of finding fault. This meeting commented that because the mark schemes must all be phrased positively 'they suggest an idea of perfection which is not attainable' (Senior Examiner, English 2). It may be a care to remain close to the mark scheme that keeps the senior examiners from demonstrating the same kind of balance as their subordinates; the Team Leader in History 2 comments on her own analysis of one script that despite 'saying all these positive things, it didn't get full marks.' The expression of the mark scheme does not necessarily encourage a balanced analysis.

Analysis outside the script

In both break-out and plenary, the senior examiners leading the training are as concerned to analyse the questions and their assessment objectives as the answers. This analysis seeks to

prevent or warn against assumptions which might be made by examiners; in some cases the overwhelming presence of some feature or other might lead examiners to believe that feature to be a requirement, and in others the trainers want the examiner to see requirements of the question which they might otherwise have missed. So in the plenary session of History 2 the Chief Examiner embarks on an extended lecture to the examiners on the second section of questions:

> there is nothing there about AO2a [context] because part b does not assess it. Part b does not assess AO2a, do not look for it, they are not required to do it.... they are required to analyse and evaluate... they are exploring a claim and in order to explore that claim they are looking at the weight of evidence for and against it.
>
> So you can see in the way that that last student did all sorts of unnecessary routine, frankly silly, comments that that student has been taught to do it. That's a misconception that you do not need to take into your marking.

The concern is clear that if an examiner sees enough candidates doing something, in this case commenting on the historical context beyond that which is given in the sources for the paper, then the examiner will come to believe that all candidates should be doing it. It is interesting that the senior examiners do not feel that relying on the mark scheme, which breaks down the requirements of each paper, is enough, and that the point must be insisted upon.

The Team Leader in History 1 makes a point about a specific question, which was seen as one of the common scripts in the morning, and of which his team of examiners have seen two further examples during the break-out session. He asks his team if they think that there might be something missing from those two answers 'something that ... really <u>top</u> candidates will have actually seen in that question?' Their attention having been drawn to the wording of the actual question, the examiners noticed that in fact it deals with the period after the Battle of Hastings rather than the battle itself – the name having provided a signal which automatically set both candidates and examiners thinking along specific lines, as if it were another 'false friend' like those considered above. Although the Team Leader then goes on to make a caveat, to ensure that if no candidates interpret the question completely correctly, then they don't 'debar people from getting a really high mark', breaking down the requirements of the questions enables the examiners to see much more

clearly if the features which they identify in their analysis of scripts are relevant to the judgement which they must make. (This episode is considered further in Chapter 7).

The other examiners also linked their analysis of scripts to analysis of questions, seeking to find a broader application than merely making a judgement on a single script. One of the examiners on History 1, for example, considered in great detail the balance of content in the answer between the two 'stated factors' of the question, and the level to which each was developed, which then led him to ask the Team Leader the more general question: 'if we have two stated factors as we have here, is development of one sufficient or do they both have to be developed?' The detailed analysis also extends to the candidate's grasp of the topic versus their grasp of the question: these examiners are very precise in their consideration of the material during the training meeting, perhaps setting up a clear understanding on which they can draw later, when at home alone.

The Language of Analysis

Much of the analysis on all papers and by all parties is couched in jargon. To some extent the entire training meeting is an exercise in giving the participants the key to that jargon, in defining and exemplifying key terms. Some of it is subject-specific lexis, with which the teachers who are training to be examiners will already be familiar, like the 'adjacency pairs' which form the basis of so much discussion for English 1. In other cases there are terms which appear to be normal phrases, like the ubiquitous 'own knowledge' which have specific applications which differ between subjects and sometimes between papers. In many cases these terms are drawn directly from the mark scheme.

On the one hand linking analysis of the script to the mark scheme would seem to be a useful exercise, but describing the characteristics of the script only in terms which are also used in the mark scheme can in fact be counter-productive. Comments such as 'a crucial point in level 2 is the material is unlikely to be developed very far - the material here is not developed very far at all' (Principal Examiner, History 1) do little to help examiners understand what, precisely, that means. Specific exemplification is much more helpful in creating a gloss of the terms in the mark scheme and allowing examiners to understand exactly what they are looking for.

One of the difficulties of the language of the mark scheme is the repeated use of individual terms with slightly differing adverbs to indicate gradation. As examiners analyse the scripts and try to reach a judgement on a script they wrestle with these gradations, so that we get one examiner commenting that a script is 'probably more implicitly than

explicitly analytical but it is consistently' (Examiner, English 2), and another in the same paper identifying a script as using 'a range of relevant terminology – whether it's a "wide range"?' (Examiner*, English 2). (This topic is considered further in the next chapter.)This difficulty of gradation as reflected by the modifiers being used is not restricted to the terms taken from the mark scheme, however. The word 'sophistication' or 'sophisticated' is repeatedly used with differing modifiers on all four modules, but appears on the mark scheme of just one. The modifiers tend to be repeated too, so that many things can be 'limited' or 'reasonable', as if they were terms which matched agreed widths of measurement on a scale.

The analysis in all four units shared terms which were not drawn from the mark scheme but which clearly conveyed a shared meaning to all examiners. The 'fluency' or 'fluidity' of the writing are two such terms, but another word was striking in its appearance in the transcripts: 'clunky'. These words, which suggest value judgements, are typical teacher terms, which though they are not defined by the meeting, nevertheless carry a shared weight borne of a community of practice: a form of guild knowledge. These everyday terms, which appear to be less defined, with less gradation, than the more specific terminology of the mark scheme, may therefore be easier to use. There is a shared understanding, which may well not be perfectly aligned, but these are terms which examiners are happy using because they reflect the value judgements of their everyday teaching lives. In some ways this reflects the theory of an underlying construct to which examiners are making reference.

Analysis in the 'think aloud' protocols

This section considers the way in which the script analysis modelled in the meetings is carried through into the three 'think aloud' protocols, which are drawn from examiners on the History 1 and English 2. The situation is a radically different one: examiners are not trying to demonstrate their understanding, and instead are most concerned about the numerical mark which they are going to assign to the script they are reading. The 'think aloud' protocol does mean that a line-by-line commentary is established, but the analysis is considerably sparser than that of the meetings. Nevertheless, elements of the modelled analysis appear in that of the examiners, and the characteristics outlined above apply too.

The two history examiners, Parrot and Caspar, both used the factual content of the essay to establish an outline of the answer which they are considering. Their ongoing remarks are therefore more commentary than they are analysis, creating a listing effect. In the course of this list, they both noted the existence or dominance of the stated factor,

although not in so many words: for them it is the 'Duma' or the 'military mistakes' that is repeated through their discussion and then weighed at the end. Thus Parrot noted 'now it talks about the Duma it's going to be at least a level 3.' Given that neither of the two used the terminology which was established during the training session (when the senior examiners needed to demonstrate the general rule, as opposed to the examiners' specific task at hand) it is likely that the insistence on 'information' and 'facts' which they both demonstrated is linked to the 'other factors' of the training meeting. At one point Parrot did say that an answer 'just doesn't mention enough about the other factors', but it is an unusual comment in terms of its expression. Occasionally Parrot also extrapolated from a specific factual detail to the more general category of factor that it represented: 'it begins with McCarthy – in other words pressures at home.' It may not be demonstrated often in her vocalisations, but it does appear that some level of analysis was occurring during the focus on factual content.

This is also supported by the fact that both history examiners noted the level of detail which a candidate had produced to substantiate the factor in their answer: 'reasonable detail there' (Caspar). However, it is Parrot who was most concerned with detail, making a comment related to it on almost every essay, and certainly for every candidate (the History 1 examiners mark both essays a candidate produces, one after the other), so that she identified the 'incredible detail' of one essay, and praised another for being 'highly detailed', while the phrases 'lacks detail' and 'lacks any sort of detail' appeared frequently in her vocalisations. There was little exemplification from either historian: by the summation at the end of the essay they tended to be operating at the level of general characteristics, such as the descriptive/analytical decision. Parrot rarely made any summative comment at all, going straight to the mark scheme to make a decision, which was not often justified with reference to examples.

Lupin, on the other hand, made very specific, detailed comments on the scripts which she reads, and undertook careful analysis to answer key questions, such as whether a candidate had taken a literary or a linguistic approach, picking out key phrases and specific examples from throughout the essay. The consideration can take more than a minute:

> there's not a lot of linguistic terminology though is there? There's not a lot of sign of a language approach - again it's more of the literary approach. There's nothing really about language is there... Yeah there's this symbolism and this staged dramatic ending but it's not language [*breathe in through teeth*]. There really isn't anything

> there so it's not integrated.... it's pathetic fallacy so we can just about – that's really literary isn't it?

She picks out specific words from the text to judge whether a candidate has used a range of terminology, or a wide range. It is interesting that Lupin's analysis tended to be much more specific and located in the essays which she was reading, given that her training meeting had not provided her with a complete model of analysis.

As with the training meetings, the amount of analysis which was carried out during the 'think aloud' recordings on very good scripts is extremely limited, and indeed the amount of commentary dropped significantly. One essay which received full marks from Caspar provoked only three comments, which still managed to contain four superlatives and the comment 'just can't fault it.' The only very bad essays in the data were among those being marked by Lupin, but unlike in the training meetings the amount of analysis was not reduced, as she double checked which level of the mark scheme they represented, repeatedly going over the script in an attempt to find more marks for scripts because of her sympathy for the candidates (which did not prevent them getting low marks, despite her remarks that it was a 'shame').

All three examiners considered the analytical/descriptive distinction, and Parrot also introduced the word 'narrative' as a third term in the same group, not as a synonym for descriptive, but as an equivalent level of skill. For the historians it was a simple unthinking judgement, and descriptive seemed to be pejorative, so that 'descriptive start but accurate' (Parrot) is a balanced comment. Lupin's discussion of the distinction was more complex. On one essay she commented: 'is it descriptive? It's more than descriptive isn't it? I think that's critical understanding' and then two minutes later, trying to come to a judgement she decided 'there is some analysis there, yeah.' It seems that rather than being a two-way distinction, in which it is either descriptive or it is not, there is a sliding scale incorporating critical understanding along the way.

All of the examiners paid attention to the stimuli which were suggested as being signals of quality by the analysis of scripts during the training meeting. The existence of a plan, without even considering its contents, is something which all the examiners note; Lupin greeted any plan with the word 'promising' and was pleased to see a 'complex plan'. The very good essay which Parrot marks has such a good plan that it sparks a longer commentary than the rest of the essay:

> I'm expecting a level 5 looking at this plan... in the plan it's a level 5 already... oh the plan continues onto the other page... incredible...

> highly detailed...also the negatives [i.e. it also considers them]... I can see from the plan this is going to be top level 5 if it's well written.

The introductions and the indications which they give about the direction of the essay were frequently mentioned: Caspar mentioned approvingly the 'clear line' demonstrated in the introduction. The existence of a conclusion, something which was mentioned a great deal in the plenary session of History 1, is important for all examiners, in terms of its relation to the argument, and in fact Parrot goes back to reassess one essay because its conclusion contradicts the impression she had formed from the rest of the essay. Key words are a particular stimulus for the English paper, as identification by the candidate of specific linguistic or literary devices signals a particular level of skill.

The importance which the history examiners place on factual content has already been noted, but for Lupin it is the 'approach' which is the key aspect. The English 2 mark scheme cites an 'integrated approach' as a characteristic of a top level answer, which is to say an essay which combines literary and linguistic analysis and uses one to support the other. For most essays she considers whether the approach is one or the other, or – rarely – integrated. Another key signal for her, which is not relevant to the history modules, was the length and amount of quotation which a candidate uses, so she comments 'oh dear, it's too much quotation' a number of times, which reflects the attention paid to whether quotations were embedded or not during the training meetings for both English modules.

This element of essay style is related to the question of Quality of Written Communication, which is essentially a completely marginal matter for Lupin, who makes very infrequent reference to how well-written something is, with the potential exception of one essay where the candidate ran out of time and switched to bullet points rather than continuous prose, to which she objected. For Caspar it is also marginal, although comments about the fluency of the writing occur a few times over the course of the 'think aloud' data. Parrot, however, frequently refers to the writing style and particularly uses it as a decider between two marks, commenting, for example, 'I'll give it 15 though because it's quite well written.' For her content and the Quality of the Written Communication appear to be the main focus of her analysis.

Just as in the break-out sessions, the analysis of all three examiners tended towards the balanced, with a mixture of good and bad comments. In the case of Lupin in particular, this balance was the result of going back and forth to decide on a mark, picking out examples which support going one way or the other. Parrot and Caspar tended to express

the balance in the body of their commentary, so that most comments have a 'but' or a 'though' included:

> Now we get onto description straight away but good intro. (Parrot)
>
> Gives a lot of detail. Doesn't link it to the question though. (Caspar)

The balance is, I suggest again, the result of the fact that the examiners have to weigh up the different elements as established by their analysis in order to reach their judgement, while the senior examiners' analysis is instead aimed at demonstrating a route to a particular mark.

Conclusions

The analysis of the scripts in all modules can broadly be divided into four categories which demonstrate the foci of attention which are being modelled by the Principal Examiner's script analysis, and which are reflected in the break-out groups. The first is the content of the script, that is to say the factual elements within it, such as the historical factors in History 1, or the types of literary and linguistic features identified in the English papers and the presence of technical terminology. This is simple identification of the presence of information, and when necessary, a judgement as to its accuracy.

The second and third are closely related and require deeper consideration on the part of examiners: they establish both *what* a candidate is doing and *how well* he or she is doing it. This is the separation between, for example, description and analysis, which is a key distinction in both of the subjects under consideration here, and one which is made in all four modules. Then, if a candidate's response is in the domain of analysis rather than description, a judgement as to how well they are analysing is contained in the gradation of the modifiers used to describe it. They may be 'attempting' analysis (a term used in all four papers), they may be 'fairly analytical' (Team Leader, History 2) or they may be accomplishing analysis 'with some success' (Principal Examiner 2, History 2, but it's a typical remark). 'Undeveloped' is a key word which is used to establish lack of success in many areas. These two foci relate most closely to the mark schemes, which are likely to distinguish between description and analysis, or assertion and argument, as a marker of the upper or lower halves, and then to use gradations of success to place scripts within that half, or indeed within a level, to a specific mark.

The fourth focus is one which could be folded into the category of what a candidate does, but which is differentiated enough to warrant its own discussion. In all modules, at all levels of the hierarchy, examiners mention the 'approach' which a candidate has taken. This

can be, and often is, a reference to the structure of the response, for example in History 2, the approach is usually dichotomous between grouping the sources and dealing with them discretely, in order. This is an example of how 'approach' comes close to the second category. It can be more of an attitude or an ethos, however, which is used to discuss and make concrete the feeling engendered by an essay; so in English 2 the examiners contrast a 'checklist approach' with one which is more analytical. In some cases discussion of the 'approach' conceals deductions by the examiners of how a candidate has been taught, or what they were thinking when they composed their answer; 'it's more like a media studies answer' (Senior Examiner, English 2). The inclusion of this nebulous concept, in an analysis which one might think should reduce the element of subjective judgement by identifying specific characteristics so that they may be matched to the mark scheme, may seem odd. However, in some ways having a non-specific, undefined concept allows the discussion of things which might not be easily construed in the terms of the mark scheme but which the examiners feel should be valued. That the mark scheme does not enable examiners to deal with all the aspects of a script is evident from the plaintive comment by the Chief Examiner on English 2 that a particular script 'does relish the text and it would be a pity not to reward that. We used to give marks for enjoyment for literature' (an attitude which recalls the debate over whether analytical criteria can capture all aspects of a qualitative judgement).

To return to the definition of analysis raised in the introduction to this chapter, it can be seen that examiners do indeed use the process to reduce essays to different elements, and to identify those elements, sometimes with detailed thought, sometimes instinctively or potentially without good reason, which enables them to create a simplified 'map' of the answers to aid them in reaching a decision. In some ways this corresponds to the 'visual map' that Crisp and Johnson (2007) suggested was created by annotation, which helped to reduce the cognitive demand of marking. The annotation of scripts is much harder using a computer-based marking programme with scans of essays, but the 'think aloud' process potentially allows examiners to create a conceptual, oral map of the answer to the same end.

Analysis of scripts is ultimately a tool which Principal Examiners and examiners put to different uses, and which enables the researcher to see some of the foci of their attention during the marking process, and some of the factors which influence the judgement they make. Those factors include both those mandated by the mark scheme and additional concepts such as the 'sophistication' of the answer; the terminology is drawn from the mark scheme and from a shared vocabulary of teaching. Individual features of scripts can be

considered on their own merits or in terms of the more general category which they represent. The presence of some features can also be treated as a signal of a certain quality of answer; these signals attract the attention of examiners, although their supervisors warn against too simple reliance upon them.

Chapter 6: The role of the mark scheme

"*Tis with our Judgments as our Watches, none go just alike, yet each believes his own." Alexander Pope*

The mark scheme holds a central position in the marking process, and is designed to hold the absolute standard against which all marking judgements are measured. In essay subjects such as English and History the mark scheme is composed of generic descriptors of levels of attainment and a passage of question-specific 'indicative content' (although not all boards or specifications include this, all units in this study did). The generic descriptors remain the same from year to year; they are widely available via the internet to teachers and are often used in schools, so that candidates are likely to know of their contents. Some may make an effort to adjust their writing accordingly. Indicative content, used widely in the host organisation's specifications, is neither exhaustive nor a list of information required for an answer to receive full marks. It is simply a nod to the fact that examiners may be marking examinations on topics within their subjects with which they are not familiar. One team leader suggested that if markers found an element in an answer that was not in the indicative content and that they did not know themselves, they should 'Google it' (H2). It seems likely that this is a widespread practice. The importance of the indicative content is suggested by a Principal Examiner on History 1 who warned his markers to make sure they had studied it, on the basis that in the previous session 'examiners who were not as familiar with certain topics as others tended to be more generous within the levels.'

In the assumed model of examination marking the mark scheme (including the indicative content) is considered side by side with each essay, and the generic mark scheme drives the process of judgement. It has already been mentioned that the stated purpose of the standardisation meeting is to ensure that examiners have a 'common well-founded understanding of the mark scheme', yet there is a lack of explicit focus on the mark scheme as a document, or on understanding it *per se*, as opposed to seeing how it has been operationalised by senior markers. The mark scheme is generic and it is abstract; much of the activity of the training meeting is devoted to seeing how it can be made concrete. The approach is summarised by a sentence from the common briefing: by the end of the meeting 'you will have marked sufficient responses to familiarise yourself with the mark scheme.'

All meetings used both the generic mark scheme and the indicative content to varying levels; as meetings progressed and discussion moved farther from the lead of the

Principal Examiners, the number of mentions of or quotations from the mark scheme dropped. The exception was the English 2 meeting, in which examiners continued to refer closely to the mark scheme throughout, debating terms and application in thorough consideration of the scripts to hand, although as I have mentioned, that meeting was dominated by senior examiners for its duration, and never gained any distance from the Principal Examiner. Even English 1, led in plenary by the Principal Examiner throughout, became less methodical and closely-referenced to the mark scheme as the meeting went on. This Principal Examiner, and the vast majority of other trainers, both principal examiners and team leaders, enabled examiners to take short cuts through the mark scheme by providing rules of thumb and suggesting shibboleths which could be used to make decisions about scripts. Sometimes examiners themselves would propose them, particularly experienced ones, and seek approval from their supervisors. For example, an examiner on History 1 in the break-out meeting suggested:

> one of the points is that, as, as long as they're attempting any analysis at all, that's the crucial point to get to level 3 isn't it? If there's any analysis attempted it's going to be a level 3 answer. It's very much a sort of turning point.

This key criterion then provides a hurdle over which a script must "jump", and a simple rule that is easy for examiners to remember. Other circumventions of the mark scheme are dealt with in Chapter 8, which looks at heuristics, both official and unofficial. Instead this chapter is concerned with the way the mark scheme, including the indicative content, is presented and used during training and marking.

Understanding the mark scheme

The mark scheme represents another level of difficulty in the marking process in its use of modifiers to indicate levels of skill, briefly discussed in the previous chapter, so that examiners must know the distinction between, for example, 'mainly accurate', 'generally accurate' and 'accurate'. This is something which can only be done by exemplification, or by understanding an underlying construct; there is no way of explaining verbally what the distinction would be. There was considerable discussion on these types of points in English 2, where the mark scheme had never been used before, where an examiner might raise a particular phrase from the mark scheme: 'this "uses a range of relevant terminology". Whether it's a "wide range"...' The rest of the team would then judge the script according to that characteristic, in the process coming to an agreement on what constituted a 'wide

range of relevant terminology.' The nuances of the adverbs used in the mark scheme and their prominence were not necessarily alien to the examiners, as one, again in English 2, commented that a script was 'probably more implicitly than explicitly analytical but it is "consistently"'. Later the Principal Examiner tells the team 'I'm still looking for that "precisely analytical"', and none of them query the shades of meaning in these terms. Sometimes, however, they do cause difficulty: the Principal Examiner of English 2 on script describes himself as 'sort of hovering between 8 and 9 [marks] because it says "significant range of literary and linguistic" [terminology].' The meaning of the word 'significant' should be as yet undetermined in this first examination of this specification; he clarifies his uncertainty with the words 'perhaps there could be more terminology' but there is no easy way to define significance in this context. This group of examiners had more discussion focussing on the meaning of the mark scheme than any other; 'I think that the problem is the wording in the *critical* understanding' says one examiner. The focus on the exact expression of the mark scheme and its meaning is probably at least partly due to the fact that it was the first time that paper had been examined. Despite the mark scheme being composed of familiar components, there was no bank of previously examined scripts to form a framework against which to consider new essays.

Some terms in the mark scheme are unproblematic: no-one would question what it would mean for a script to 'compare', for example, but there are other terms which require greater explanation. At various points during the standardising meeting senior examiners, when considering the mark scheme, would use words or phrases that were not taken from the mark scheme, for example one Principal Examiner suggested that when deciding on which mark within a level a script should receive, examiners

> might consider range and the selection of material. Whether the argument is sustained or not, um, quality of written communication. Is there a chronological coverage of the period? (PE1, H1)

This alternative to the criteria provided in the documentation gives the examiner a set of characteristics which are not graduated in the same way as those written in the mark scheme but appeal to professional expertise. The discriminating factor suggested in the mark scheme to decide on a mark within this band is the script's being 'convincing in its range and depth' of material. This can be seen within the first line of the Principal Examiner's comment but the other phrases are not reflected in the mark scheme, and are not providing fuller meaning to the 'range and depth' cited in it. Providing alternatives to the criteria was a

frequent strategy; glossing of individual words was not seen in relation to the mark scheme, although it was seen elsewhere to explain jargon terms and acronyms.

Sometimes examiners asked explicitly for clarification of a term in the mark scheme; they never received a response that constituted a clear definition or definitive explanation. The example of the unhelpful response to 'what would be analysis?' by the Team Leader in History 2 was given in the previous chapter:

> **Team Leader**: It may be that they're analysing up here [*indicates lines*]. It may be that they've got their argument and they're just grabbing - they wouldn't be the first historian to do that.

Although the entire process relies on the use of examples to inculcate an understanding of what each level's criteria 'mean', this particular exemplification would not be of much use in gaining a definite knowledge of what constitutes 'analysis' in the context of the A level history examination. It may also suggest that even senior examiners' understanding of terms is in fact more instinctive and more reliant on an overall 'construct' referenced approach since a precise explanation cannot be supplied. However, it may also be akin to a feature identified and discussed in Chapter 8, that of 'productive uncertainty' where trainers refuse to give categorical instructions to examiners so that they do not over-rely on a given definition or instruction, thus leading to bias. A narrow definition of what constituted analysis could only be given by laying down a set of characteristics which, like the indicative content, ought not to be considered either exhaustive or necessarily present for analysis, but would undoubtedly be misused by some. Finally one might think that an historian would be able to identify 'analysis' given that it is such a key skill in their own expertise.

The mark scheme during the training meeting

Examiners keep the mark scheme to hand throughout the meeting, sometimes making their work spaces unwieldy with paper, given the number of documents to which they must keep referring. In a physical training meeting this is particularly problematic given the limited space on the table which the team of examiners are gathered around; even when at home marking on screen, they must still keep at least the multiple pages of the mark scheme and indicative content on their work surface. Examiners are explicitly reminded of the various papers they will require in the history meetings ('you'll need the paper, the sources and the mark scheme handy'; Principal Examiner 1, History 2), but it is also, one assumes, true of the English papers. While attendees at the physical meetings are given packs of print outs and

photocopies, those with online training must download and print their reference material. Some discussion at the English 1 meeting did suggest that not all examiners had done this beforehand, although their briefing material instructed them to.

One examiner did demonstrate regulated rule-based application of the mark scheme during the training meetings, but she was the exception rather than the rule. Examiner 7 in the History 2 module explained one mark she had awarded in the following way:

> It's well-integrated. It's Level 3. I didn't give it a level 4 (.) I usually work down like this (.) cos the reasoning (.) the way they use the sources it's not developed… it's integrated… it's weak but there's a discussion and the sources are still there from beginning to end.

The adjectives 'developed' and 'integrated' are taken from the level descriptors, and she suggests that her normal procedure is this systematic working through the levels, which might be considered as an ideal practice by the Awarding Body. Such close and methodical adherence to the mark scheme is not demonstrated anywhere else in the data from the training meetings. Nor does she show the same response when she speaks at other points in the meeting.

The indicative content remained virtually unused during the meetings. Examiners preferred to ask one another if a factual detail was correct rather than to consult the indicative content:

> **Examiner:** can someone answer me a question? Was Othello born a slave - have I been misreading the play for 30 years?
>
> **Chair of Examiners**: Othello was nobly born. It's made a point of in the play.

The entire debate on the expected content of the Battle of Hastings essay which caused the History 1 Team Leader's extended passage of questioning, which was discussed in Chapter 4, was conducted without a single reference to the indicative content by any of the examiners involved. A single mistake in the indicative content was highlighted by one of the Principal Examiner for History 1; on the whole it seemed to be either an unnecessary safety net or a stimulus too far for examiners to spare it any of their already stretched attention.

Introducing the mark scheme

The introduction of the mark scheme was discussed briefly under structural aspects of the training meeting in a previous chapter, where it was suggested that it was surprising that a

specific and explicit introduction to the mark scheme was not a mandated part of the standardisation meeting. Only one module, History 2, demonstrated an explicit focus on the generic mark scheme during the training meeting, and this appeared to be an afterthought on the part of the Chief Examiner.

It might have been expected that the meeting for English 2 would spend time examining the mark scheme in some detail, given that it was the first time the module had been examined. In fact, although all the participants of the meeting had the mark scheme in front of them, and referred to it throughout, the business of the meeting was to launch directly into reading a script and for examiners to decide on a mark for it, with a simple reminder to deal with the two Assessment Objectives separately being the only mention of the mark scheme. In some ways this is appropriate to the proportion of senior examiners in the room, but it is still surprising. The lack of introduction was striking given the much greater prominence of the mark scheme (compared to the other modules) throughout the meeting.

Although there was no generic introduction to the mark scheme at the beginning of the History 1 meeting, the first Principal Examiner for History 1 *did* provide for the examiners an extremely abbreviated description of each level before they had to make their first decision on the level a script should receive:

> Level 1 is of course detached and disconnected simple statements. Level 2 is simple statements that are making, er, a relevant point. Level three is where candidates are beginning to explain. Level 4 is where candidates are analysing, and level 5 is where they're analysing evaluating and, er, show a wide range of good qualities.

These summaries reflect but do not replicate the first sentences of each level descriptor provided in the mark scheme for this paper, which are given in the table overleaf. Words which appear in both summary and mark scheme are in bold.

Level 1	Candidates will produce mostly **simple statements**.
Level 2	Candidates will produce a series of **simple statements** supported by some accurate and **relevant** factual material.
Level 3	Candidates' answers will attempt analysis and will show some understanding of the focus of the question.
Level 4	Candidates offer an **analytical** response which relates well to the focus of the question and which shows some understanding of the key issues contained in it.
Level 5	Candidates offer an **analytical** response which directly addresses the focus of the question and which demonstrates explicit understanding of the key issues contained in it.

Although these summaries of the level descriptors are rough and ready, intended to inform an instant judgement, rather than the wider marking period, they do offer a much shorter and more memorable précis of the mark scheme, which may provide a hook for the examiners to engage with it. They use potential synonyms, and in some cases actually add detail; such as level 1's 'disconnected' statements. The 'wide range of good qualities' for level 5 is very non-specific, however; it relies on examiners' professional judgement both as to what a 'good quality' is, and what constitutes a 'wide range'. Just as with the introduction offered to the mark scheme for History 2, these summaries do not paraphrase or define, for example, what 'analysis' is. This key term is in fact found at three levels of the mark scheme; the Principal Examiner sidesteps this by using a different word – 'explain' – for the lowest level of analysis, and then adding 'evaluating' to it for the top level. This differentiation was seen more clearly when the Principal Examiner repeated this paraphrasing procedure at the end of the second training script, with an even more abbreviated form of the level descriptors:

> is it level 1 - simple and modest? Level 2? Simple statements? Level three: an explanatory focus, level 4 an analytical focus, or level 5 an evaluative approach? (PE1, H1)

This is the final time that he does this: after the second script examiners are assumed to have grasped the fundamental descriptors for each level, and to be able to make the decision without the prompt. Thus it seems that this Principal Examiner's approach to training examiners to the mark scheme is to abbreviate and repeat, thus circumventing the difficulty of committing the entire set of descriptors to memory. Examiners also of course have the paper or pdf copies of the mark scheme to hand, and the ability to refer to them,

but a mental conception of each level may enable them to participate more easily in discussion and to make the quick decisions necessary during the training meeting.

Another approach is taken by the Principal Examiner on English 1, who has the advantage that her examiners 'all know it very well by now', as she tells them. She only reads the descriptor of the band into which the current essay falls, rather than providing an overall consideration of the mark scheme, interpolating some comments on how the essay meets those descriptors as she goes:

> We need to look at the descriptors to work out, to shift within the band and to decide where we're going to place, um, this particular script. "Develop comments on context" - it does. "Responses include well developed links between the language of the text and the context in which they are either produced or received. Examines both extracts: at the bottom of the band detail across the extracts will be consistent and thorough; at the top of the band there will be some evidence of *sophistication*."

She stresses the final word to indicate it will be a key criterion for placing this answer, and in earlier discussion of the answer she had described it as 'seductively sophisticated', echoing the language of the mark scheme; her final mark is indeed at the top of the band. The number of comments she adds to the descriptors varies; she tends to add more when the script does not merit the complete positive application of the descriptor. For example, after beginning 'it does offer "critical analysis", it does "identify links between form and function", um but the "specific analysis" is somewhat restricted and limited, she continues for a further minute considering how the essay does not precisely meet the other descriptors.

It can be seen therefore that the mark scheme is introduced not as a central aspect of the training meeting, but rather as a tool, to be used to make the correct judgement as to the location of the sample scripts. While the nominal and stated purpose of the meeting is to understand the mark scheme, the business of the meeting is entirely focused on making judgements, and the range of tools that can be used for that, of which the mark scheme is just one.

Checks and balances

During the plenary sessions the Principal Examiners frequently demonstrate the use of the mark scheme as a check on a decision which has been made, or to fine tune a decision

within the broad judgement. So for example one Principal Examiner, having identified his first script as a top level (5) embarks on this discussion:

> the mark scheme says in level 5 candidates are going to certainly analyse. They will analyse er the given factor or factors or whatever it is in the question, they directly address the focus of the question, which this answer does. They demonstrate explicit understanding of key issues which the answer does. The answer is broadly balanced in treating key issues, and is supported by accurate relevant and appropriately selected material, demonstrating some range and depth. (PE1, H1)

Having established that a level 5 is justified, he moves on to deciding on a specific mark, which involves drawing out the characteristics of the answer, not in the terms of the mark scheme, but more generally. He decides to move downwards because 'the answer is not particularly strong in some areas' and having done so uses the level descriptors for band 4 as a check that it has been placed correctly:

> if you check that against the level 4 descriptor: analyses and shows some understanding, supported by accurate factual material mostly relevant. Well all of the material here is relevant. Selection of material may lack balance - well it doesn't really lack in balance, and so it does get to level 5 in the lowest point within level 5. (PE1, H1)

It is noticeable, however, that although the Principal Examiner links descriptors to the script, he does not provide evidence of how it fulfils these descriptions: exemplification is general, rather than specific. Nor does his commentary during the reading of the essay refer to the mark scheme. He demonstrates the use of the mark scheme in a quite systematic way, first checking the chosen level, and then the one below (and potentially, one assumes, the one above if it were not a top level answer). He does not demonstrate this type of checking on all the scripts, but it is shown a number of times in this unit and in others. Similarly the Principal Examiner on English 1 used the descriptors in the mark scheme, but only after her examiners had selected the correct band. By checking the answer against the characteristics in the mark scheme, she framed the decision on the mark within the band within the terms of the rubric. The examiners also are encouraged to use the mark scheme as a means of checking their thinking; the phrase 'go back to' the paper, assessment objectives or mark scheme was often used by trainers to suggest this.

Although the English 2 meeting used the language of the mark scheme a great deal, as will be discussed in the next section, the generic characteristics outlined in the criteria were offset against more question-specific features, or characteristics which did not appear in the mark scheme. The final sample script, for example, prompts the following discussion about what mark it deserves in the Assessment Objective focussing on contextual knowledge:

Examiner *:	is it band 5?
Examiner:	it certainly shows understanding of context
Principal Examiner:	it's more than *some* understanding
Examiner:	you just want some awareness of the different times that the plays are a) written and b) set - I think I'd expect that in the higher bands.

The generic phrase 'understanding of context', which is modified by 'some' for the lower level, is translated into a precise expectation by an examiner, though still one which is relatively general. (It is interesting that this is in contrast with the indicative content which is not divided into levels of difficulty.) Thus the mark scheme is balanced against other concepts of quality which depend on related but unspecified things.

Talking schematically

All participants, consciously or not, echo the terms of the mark scheme in their talk. They may lift phrases from the rubric or produce paraphrases that mimic it closely, when discussing papers. Jargon terms lifted from the rubric, such as the 'debate' or the 'stated factor', the 'form and function', are particularly frequently used but they are not the only ones.

Most examiners demonstrated this trait in the phase of the training where they were required to comment on scripts which had just been read, or to justify the mark which they had just awarded. There was a tendency towards shorter explanations rather than extended ones, but both were likely to have the same amount of content lifted from the mark scheme; it was as if an abbreviated lift of words was considered to be enough justification in itself. (This is not to say that all explanations took words from the mark scheme, but many did.) Thus, for example, in History 2, the words 'provenance', 'cross-

referencing' and 'own knowledge' frequently crop up, as justification for a script either reaching a band or not.

The exception was in English 1, where the way in which the discussion was structured by the Principal Examiner meant that examiners did not analyse the scripts, instead just giving levels and marks. On occasion one might refer to an element of the mark scheme to justify the mark they were awarding, often referring to the amount of analysis which had been done, but it was infrequent.

In English 2, the vocabulary of all examiners frequently referenced the mark scheme in their discussion of scripts, often weighting one aspect against another ('they know a lot of terms but it doesn't actually analyse' Examiner, English 2), but their discussion was wide-ranging and it is unsurprising that many of the terms which they used also appeared on the mark scheme, given that they are drawn from the common lexis of the subject. 'Analysis', for example, appears on every one of the mark schemes of the four modules in this study in one form or another and is a major concern of both history and English specialists. Similarly 'context' was mentioned even when the Assessment Objective under discussion did not assess it. Terms which were not such staples of English teaching were not raised so much; 'attitudes and values', a key term for the mark scheme for English 2, is mentioned by name on only two occasions during the meeting.

On some occasions phrases from the mark scheme are used to reframe the discussion of scripts, refocusing on the mark scheme after a qualitative assessment of a script which, though focused on apparently relevant characteristics, was not precisely to the point. The Team Leader on History 2, for example, after an extensive assessment of the use of sources in the paper – a key assessment focus, but which has been framed informally – concludes 'but it doesn't "analyse and evaluate" sources and if you go back to the paper that's what they're supposed to do.' This is related to the checks and balances of the previous section, but it also acts as a signal to demonstrate that the assessment is based on appropriate characteristics taken from the rubric and thus 'proving' it is good marking behaviour.

The use of the phrases from the mark scheme is dual purpose. From trainers it acts as a way of demonstrating the applicability of the mark scheme, of implicitly exemplifying its terms by attaching them to the concrete examples of the sample papers. For examiners however, using the words of the mark scheme is a way in which they can demonstrate their competence in judgement and justify the decisions which they have made. When marking is carried out on physical paper scripts, as opposed to by means of scanned pages on a

computer screen, the annotations which examiners are required to make often echo the mark scheme. The reasoning which they must share during the standardisation meeting fulfils a similar purpose, in convincing their supervisors of their correct thinking. It seems as if part of the effect of the training meeting is to channel the thinking of examiners into certain terms, or at least their talking, so that their conscious decision-making may be based on the features which are nominated in the mark scheme. This use of the terminology from the mark scheme is also seen to a certain extent in the 'think aloud' recordings of individual examiners, to whose use of the mark scheme we now turn.

The use of the mark scheme in individual marking

In the examiners' 'think aloud' protocols the use of mark schemes varied; in all three cases the use of the mark scheme was most evident at the beginning of the marking period. Later explicit reference to the mark scheme is less perceptible: this can be seen as evidence for an increase in the automaticity of decision-making, or for the existence of System 1 and System 2 decisions and migration between the two. It may also be that as examiners become more confident in their decision-making they feel less need to use the mark scheme to check those decisions or to justify them. Once again the indicative content is notable for its absence in the 'think aloud' protocols of all three examiners; no apparent reference is made to it, despite Lupin's occasional debates with herself over whether a candidate has made a correct point or a simple factual error.

Lupin marks with the mark scheme as a constant reference; she is marking English 2, which is not only a new examination but is also the most complex of the units in that a single essay is examined over three Assessment Objectives. At the beginning she takes phrases from the mark scheme and assesses the essay according to the presence or absence of each, as if working through a checklist of 'context', 'critical understanding', 'terminology' *et cetera* when she has finished reading each script. She engages repeatedly with the meaning of the terms in the mark scheme, asking herself questions: 'is it integrated?' She spends a great deal of time in her first 'think aloud' session trying to identify if certain aspects of the script equate to the terms on the mark scheme.

> Is there a range of relevant contextual – very little [almost half a minute of silent thought] doesn't even use the word Victorian or – there was a mention of America... seen by society, but it doesn't actually say what society, there's no indication of time or f- there's just nothing there on context. Oh, except there was that ref—

> reference to performance which is a form of context isn't it.... "some awareness of context", yeah well that is some awareness isn't it?

Lupin has a range of ideas about what constitutes 'context' and runs through them to see if the essay can satisfy the requirement. She eventually identifies something which satisfies her as being possible to categorise as context, and measures it against the wording of the mark scheme. It is notable that she creates her own list of potential 'indicative content' however, rather than using that provided as an attachment to the mark scheme.

As Lupin progresses through her marking period, as measured by her 'think aloud' recordings, she is no longer as concerned with deciding what the terms of the mark scheme look like on the pages; she works more directly with the terms of the mark scheme in her decision-making:

> "Detailed analytical exploration and comparison, developed understanding of context, analytical evaluative approach". I don't think we're going up to "incisive and original".

She continues to use a tick list approach ('I'd have to say yep to most of that' she says, having read out the level descriptor she has chosen), but it is more quickly narrowed down to the relevant sector, instead of working through a wider range of potential descriptors from the mark scheme. She no longer needs to identify a band and then look within it: going straight to the appropriate level without comment she also stops mentioning different potential marks by her third verbal protocol recording (which sits in the middle of her marking period). That is not to say her decisions are any quicker: she continues to take an average of five minutes to come to all her decisions after finishing reading a script, but she is more confident of her summations of scripts and identification of the appropriate descriptors and marks. She is not entirely confident, even so, so that she is still debating some terms ('the whole thing is, it's not, I'm not sure you could call it original, but then what is original?') but the majority of the terms are no longer problematic.

Although in the earlier recordings Lupin would run through a preconceived list to find the 'indicative content' that was relevant to one of the key words from the mark scheme, in later recordings she no longer had to search through possibilities. While reading through the essay she would cite a list of elements, key points that had claimed her attention, and which were relevant to the Assessment Objective. Such points seemed to function as the verbal map of the answer, so that when she came to the end of an answer she could move straight to the appropriate section of the mark scheme. In particular she

tended to list aloud vocabulary in the essay that constituted 'literary and linguistic terminology'. While she is reading the scripts she does not use the words of the rubric, except for 'language' or 'terminology' which are very common terms in English, although the focus on terminology may be suggested by the mark scheme.

Caspar and Parrot, the two history examiners, do not use the mark scheme in such a close and integral way. Both use the shibboleth of the 'stated factor' repeatedly, just as the emphasis in their training taught them to, although Parrot never refers to it by that name, instead referring to the actual historical factor.

Parrot virtually never refers to the mark scheme; her focus is almost continually on the script under consideration, and its contents. She very occasionally – perhaps one in ten essays – decides that she needs to refer to her 'notes'. This is accompanied by the shifting of paper but it is not clear if by 'notes' she means the mark scheme, indicative content or handwritten notes from the training day. The first time is in the context of her assessment of the stated factor (the Duma): 'it mentions the Duma a few times in passing, but whether that is really enough?' Later however, she consults her 'notes' when unable to decide between a high level 3 and a low level 4:

> Low 4 but very low 4 or a high 3 [paper rustles]. Sometimes there's
> no gap between them but of course if you look at the notes...

This is typical of the kind of instance in which she consults the notes, that is, when encountering some difficulty in reaching a decision (her selection of a mark is usually quick and seamless). The borderlining aspect of this particular comment suggests that it is the mark scheme which she is consulting. Interestingly, Parrot is extremely concerned with 'information' and 'detail', which are the two most approving terms she has for an essay, but she never makes any reference to the indicative content, or any source of authority other than herself. On one single occasion in her recordings she makes reference to having 'just double checked the mark scheme', in a similar instance to the first one noted above, where the essay is confined to a high level 3 'because it doesn't say enough about the other factors', the aspect which most concerns her. She rarely makes any reference to the terms which are used in the mark scheme, such as analysis, although she does frequently use the word 'descriptive' which appears in the band 3 level descriptors. Since the majority of the essays Parrot marked during her 'think aloud' protocols were level 3, it is difficult to establish if she was using descriptive as a technical term from the mark scheme or simply as an appropriate adjective. It may, however, have been a key feature in determining the band

for her; she found placing essays in band 3 to be easy, judging by the comparative speed with which she did so, in a process which was already fast.

Caspar demonstrates a mix between the two approaches, although more closely aligned to Parrot's approach. He does refer to the 'stated factor' in assessing the essays, and also notes the presence of 'other factors', which Parrot only does when quoting directly from the essay she is reading. Like her, he uses the recommended technique of deciding on a band and then moving to a mark within it, but is more likely to consider the level descriptor between the two stages: Parrot simply identifies a band and then a mark in virtually the same breath. Caspar was more likely to check the mark scheme if a new essay was placed in a band that differed from the previous one.

It is clear that all three examiners use the mark scheme in differing ways and to differing extents during their marking. All had it to hand during the on-screen marking. None made any reference to the indicative content portion of the document, even when they were most concerned with content. Indeed even when they commented 'is that correct?' as they all did, at one point or another during their marking, the examiners showed no inclination to either check the indicative material or to consult the internet, as had been suggested during their training.

Conclusions

The use of the mark scheme is as varied and changing as the training methods used in the meeting; different modules use it in different ways and within those modules different examiners use it to a greater or lesser extent. The indicative content appears to be universally ignored, which may suggest that if it has any use at all, it is at the beginning of the process, in which examiners familiarise themselves with the questions which candidates have answered. It might be of more use in a summer session, when experienced examiners taking part are fewer. All examiners use the language of the mark scheme at some point, often to justify their decisions, as a validation of their own judgement; since the terms of the mark scheme are often taken from the subject being examined, as part of establishing valid assessment, it is not unexpected to see them being used in discussion.

Some examiners use, or claim to use, the mark scheme in a methodical way, applying it just as the system assumes. Many use it, as Principal Examiners seem to, as a tool to make a check on their marking decisions, particularly during the training phase. Others refer to it only when decisions become problematic, perhaps in the nature of a last resort; deliberation over marks is associated with decreased accuracy of decisions in Crisp (2008a).

Does engaging with the mark scheme at such a juncture in fact deter examiners from the use of their inner conception of what a level 4 'looks like'? Alternatively one could argue that it is on scripts which are hardest to categorise that we should expect the greatest lack of marking accuracy, and the use of the mark scheme merely signals the difficulty of the decision.

It is clear, however, that the mark scheme is not unproblematic: it causes discussion in all meetings, and there is a lack of understanding of what its terms mean precisely demonstrated by various people throughout the data. If it were easy to apply, there would be no need to employ experienced teachers to be examiners. It is perhaps overstating the case to suggest that the fact that teachers are employed means that to some extent the system *must* rely on professional judgement and a 'construct' of quality on which they mark scheme rests, but it is a possibility.

Senior examiners attempt to make the mark scheme into a more manageable tool for examiners, sometimes by reducing it to its simplest factors and sometimes by exemplifying its use. There is usually no formal introduction to the mark scheme during the training meeting, The experience of the examiners taking part has been commented on before, and is also acknowledged during the meeting, and this might be seen as a reason for the omission. However, there are a number of points where Principal Examiners say things which can have no application to these 'good' examiners, simply because it is part of the usual practice, which suggests an introduction to the mark scheme is not given as part of the normal routine at any time.

This is aligned with the slight incongruence between the stated aim of the meeting, to reach an understanding of the mark scheme, and its main business, which is entirely focused on marking. The aim of the meeting would seem, to an observer, to be learning to make judgements, or to align individual judgments to the required standard. While the mark scheme may be formally conceptualised as the absolute standard, in practice it is merely one of the tools which examiners use to standardise their marking with that of the Principal Examiner.

Chapter 7: Comparison: a principle of decision-making

> *"I believe it was Plato who said that good judgement consists equally in seeing the differences between things that are similar and the similarities between things that are different." Brian Magee, Confessions of a Philosopher*

According to Donald Laming, judgement is always relative, always dependent upon comparison (2004). The principle of heuristics also relies, at heart, upon comparison, as one can only judge representativeness, for example, against other examples. Direct comparison of scripts to produce rank ordering has been in use for checking the relative difficulty of qualifications offered by different syllabuses and exam boards since 1996 (Pollitt, 2010) and in ensuring the maintenance of standards in the same qualification over time (Bramley, 2007), and has been suggested as a possible alternative to numerical marking, an approach mainly championed by Alistair Pollitt (Pollitt and Elliott, 2003; Pollitt 2004). It is not surprising, therefore, that comparison is a recurrent feature of the way in which examiners make decisions, both explicitly and implicitly, and using a variety of touchstones.

This chapter divides the topic of comparison in examiner decision-making using three different dichotomies:

- Within scripts/ between scripts
- Primary decision-making/ confirmatory
- Rule-based comparison / unthinking

Within scripts/ between scripts[1]

Within script

Two types of within-candidate comparison can be seen in the data: between two or more answers given by the candidate; or between traits, as represented by Assessment Objectives, within the same answer.

Either of these comparisons can present a problem for the purist, as both threaten to stimulate the 'halo effect' (Thorndike, 1920), whereby perception of one trait alters the

[1] 'Scripts' is a term which can mean more than one thing in the context of examination, as discussed in the Introduction to this thesis. It can be used variously to describe both the entire script a candidate produces in response to one examination, i.e. two or more answers, or it can refer to a single essay answer. This ambiguity is reflected in the discussion here.

perception of another trait, which is a well-acknowledged source of bias in educational assessment, as elsewhere (Pike, 1999). In some syllabuses Awarding Body have taken steps to minimise this effect through the use of technology: since all scripts are scanned and distributed electronically, the two essays on one paper can be separated and sent to different examiners, and even if the same examiner marks them, they will not know the two answers come from the same candidate. However, this was not practice on all the units under consideration in this study, although all came from the same Awarding Body: it depends on the subject. In some examinations, examiners have a small number of marks at their disposal to compensate between answers, if a candidate uses material which addresses an Assessment Objective (AO), such as context, for example, in the essay where that AO is not examined, but fails to replicate that material where it would gain them the mark, in a section where that AO is examined. This is not the case in any of the units in this study.

With candidate's other answer

The practice of comparing the marks awarded to each of the answers given by a candidate on the same script was considered by three training meetings, and between them they exemplify the two different practices which occur.

In English 2, the answers are separated and sent separately to examiners, so that the marks given to each are not affected by the grades given on the other. The possibility of bias, and the necessity of this practice, is implicitly accepted but also regretted, as demonstrated by this interchange between the Chair Examiner and the Principal Examiner:

Chair of Examiners:	because I'm such a dinosaur, I find it very frustrating not to see the second answer
Principal Examiner:	we have to mark this discretely
Chair of Examiners:	absolutely

This exchange comes in the context of discussing a candidate and whether they were merely lacking in terminology or in something more fundamental; I take the Chief Examiner's comment to mean that he felt he would be able to make a more holistic (or 'fairer'?) judgement if he were able to see the second question. The use of the word 'dinosaur', however, makes its own value judgement on this, and it is clear that they both see the danger in having both of a candidate's answers available to the marker.

This is not a danger which is even considered in the History units under consideration; in fact, both training meetings used comparison between the two answers

made by a candidate as a tool. In History 1 there was explicit suggestion that comparing the levels achieved on the two essays was a useful confirmatory measure. An extended passage of the opening briefing by the second Principal Examiner focused on using this, in the context of the experience of the prior examination session:

> Something else to consider is that this time you're marking both of the items together. If you've awarded marks for both answers which are substantially different from each other it is worth reflecting on your decisions before moving on. In the summer in general, it became obvious that most candidates tend to approach both answers at the same level of um answer and therefore if you've, if you do have an answer which um has a level five or a level three for example then it's worth reflecting on your decisions before moving on. If it's obvious that the candidate has mismanaged their time then that's a different um that's a different matter um, but it is worth looking um looking at the overall script and the overall script and the overall levels. (PE2, H1)

This is another of the unofficial heuristics which are provided at various stages during the training meetings: in this case the directive to examiners to 'reflect' on the decisions which they have made by comparing the two marks awarded. A potential explanation for discrepancy is also offered, as expected, since there are no hard and fast rules to making quick decisions to be found at training meetings. This heuristic is confirmed twice more during the plenary section of the meeting, once by each of the Principal Examiners who, when going through sample scripts, point out that the other essay on that script is working at the same level as the sample.

The emphasis on this point is reflected by the behaviour of Parrot, who when marking this paper would compare the two essay marks, and sometimes adjust one as a result, considering the relative merits of the two answers even within levels, rather than just between them, so that after giving a second essay 13, she adjusts the first essay from 16 to 15.

In History 2 the comparison is both less direct and less directed. The Team Leader of the group reflects on the total mark awarded to a script to see if it is an overall fail, considering the consequences of awarding a particular mark, but does not suggest to her team that they do so as a rule. The heuristic of the same level is not suggested in this unit of

the syllabus, presumably because it was not discovered to be true in the summer examination session, and indeed, a counter-example is noted by the Team Leader:

> It's another one where you've got a stronger second one than first one - and that's often the case because they reverse it round time wise

The tendency is noted, although in a clearly descriptive rather than prescriptive way, and a potential explanation is again given. It also potentially implies that one would normally expect the first to be the stronger if candidates don't reverse their time.

Within script, between Assessment Objectives.

The other within-candidate potential comparison is between the marks awarded for different Assessment Objectives. The essay questions in History 2 and English 2 divided their marks between up to three different Assessment Objectives, which are intended to reflect different traits in the relevant subject domain, and therefore allow for potentially different performance in those traits, a classic scenario for a potential halo effect.

This potential is noted, even if not explicitly. During the plenary briefing for History 2, the Principal Examiner went through three scripts with the examiners, and for each he noted the relative performance in the two Assessment Objectives; the three papers covered the full range of possibilities, from very different, through different but 'not poles apart', to being the same level for each Assessment Objective. As might be expected, therefore, in the break-out session of this paper, the Team Leader did not compare the marks awarded for the two Assessment Objectives, only bringing them up on one occasion to point out that one candidate could not be rewarded for the same thing in both.

The essays on English 2 were marked over three Assessment Objectives, and this necessitated the most explicit comparison between them, and the most overt juggling of marks to ensure a fair overall result, which accurately reflected each trait. For example:

> **Principal Examiner:** push it [AO1] to 5?
>
> **Senior Examiner:** I think we're going to have to or we're going to penalise it on AO2 too
>
> * * *
>
> **Principal Examiner:** For AO2 are we talking about lower than AO1 –4?
>
> [general agreement]

The understanding that the answer is weaker on the second Assessment Objective affects the mark awarded for the first, in order to give them a greater range of options for the second. The use of the word 'push' also implies that if it were not for this understanding, the essay would have received a lower mark. Moments later they discuss AO4, the third relevant Assessment Objective, and the Senior Examiner comments that 'it's definitely its strongest point'. It is possible that such explicit comparison between the Assessment Objectives negates the potential of the halo effect, by encouraging examiners to rank their relative strength, rather than unconsciously follow the previous Assessment Objective. The usual method followed by the Principal Examiner during this training meeting was to ask the table for a mark for the first Assessment Objective, and then when that was established, for the next one, as a relative value. To some extent this could be seen as the *anchoring and adjustment* heuristic in operation, as the examiners find a level for the first Assessment Objective and then move up or down from there to make further decisions.

More awareness of the relative values of the Assessment Objectives is demonstrated when considering the overall mark the candidate has received; a mark which is lower overall than the group feels the candidate deserves frequently causes them to 'revisit' a mark given for an earlier Assessment Objective. This meeting had the greatest awareness that they were required to mark analytically, for each separate trait, rather than making a holistic judgement. The Chair of Examiners in particular (the one who described himself as a 'dinosaur' above) is concerned with this point, as demonstrated by his comment on one sample paper:

> just out of interest I double marked that last one. Impression marking it came out at 88% so it's harder for them to get the marks in the broken down way. More difficult for the candidates to get the summative marks. (ChaE, E2)

In fact he was wrong, as the overall mark of 56 out of 60 awarded by the analytical method is equivalent to about 93%, but it illustrated the general mood of the meeting, which preferred holistic marking to the analytical approach. Their attitude supports the challenge found in the literature (e.g. Sadler, 1989; Hamp-Lyons, 1991) that the sum of the parts of a qualitative judgement does not add up to the whole of it. The actual difference in marks (equivalent to 3 marks out of 60, which is just one mark outside the range of difference which is tolerated in monitoring junior examiners' decisions) was almost negligible, and does not contribute to the debate either way.

The comparison of the overall mark to the separate marks awarded for each Assessment Objective can be seen in two ways; either it is a useful confirmatory check on the process, ensuring that a candidate is fairly rewarded, or it biases the process which has been carefully established to ensure that the different traits of the subject domain are appropriately examined and rewarded over the syllabus. The feeling of this group of examiners was definitely the former:

Chair of Examiners: the holistic mark shouldn't run counter to your impression of the whole answer

and similarly:

Senior Examiner: I do tend to think, do I really want this to fail?

Two senior examiners, neither of whom was actually marking the paper, both emphasised to the meeting that the overall mark needed to be thought of, and that marks for individual Assessment Objectives must be considered in this context. It is clear that in the debate over the virtues of holistic marking as opposed to analytical, these examiners are most interested in validity, and are concerned that the analytical approach can reduce fairness to candidates; it suggests that they are not fully confident in the ability of the criteria to appropriately consider the aspects of an essay which a holistic judgement would, confirming the non-reducibility of a qualitative judgment (Sadler, 1989).

This consideration of the relative performance in Assessment Objectives is unsurprisingly carried through into the 'think aloud' protocols of Lupin, who marked this paper. At the point of decision-making she juggles the marks between the Assessment Objectives to ensure a fair overall mark and to weight their relative value. After considering a script's relative performance in the three Assessment Objectives, she concludes:

> That'd be 8 and 12, that'd be 20. Out of 60 [long pause]I think it's 11 [pause] Maybe it's 12 [pause] Well maybe it's 11, I think it's 11 to be honest, okay, 3 and 5 and 11. Is 19. Just checking that puts it just about T, yes that's okay.

She considers the overall mark relative to her holistic judgement, and the overall judgement of other scripts (considered below) before finally confirming the individual marks for the Assessment Objectives.

Between script comparison

There are three different potential touchstones for comparisons between scripts: the sample scripts (referred to as 'anchor' scripts in the literature); scripts by other candidates that the examiner has or is marking; and with imagined or idealised scripts for different levels.

With anchor scripts

Anchor scripts demonstrate that comparison is an expected, even a mandated, process in making judgements on examinations, although it is never explicitly stated as such in the procedures, nor in the literature on examiners' decisions. Sample scripts are provided for the very purpose of comparison, so that once training is concluded, the examiner has a set of exemplar material illustrating what represents performance at different levels. It is only illustrative, as will become apparent. The comments of the principal examiners throughout the training meetings provide interesting insight in to what is considered to be a 'good' sample script. As one might expect, all the principal examiners at some point comment on the range of scripts which they have selected, across both the range of questions and across the range of marks. The Principal Examiners on History 1 told examiners that if a question which was not represented in the anchor scripts turned out to be a common choice, then they would provide further exemplar marked material covering that question as soon as possible. One Principal Examiner on that paper laughingly apologised to the markers of one option because the range of exemplar scripts is 'slim - that's because everybody did Luther and witchcraft' (PE1, H1).

The training scripts in both history modules cover all the option 'papers' for which schools can enter candidates (which use different periods and material to cover the same skills set); as a result examiners receive a number of scripts which can only be generic examples of the standard, and therefore much harder to use to make marking decisions. Sanderson (2001) cites the limitations of the exemplar scripts provided for the markers which, unless they become so numerous that they cannot be remembered, cannot be representative of the range of marks awarded; this is exacerbated by the number of potential option papers in history, and the need to cover all within the plenary training.

Examiners find direct comparison between a script and an anchor which addresses the same question to be the most useful, presumably because it is an easier comparison to make, as remarked on directly by Lupin during her marking, and implied by the selection of reference points for discussion during the History 1 break-out meeting. There is a limit to

what anchor scripts can cover, given that training meetings happen only a few days after candidates sit the examination, and that they can only be drawn from what the principal examiners can see in that time. The Principal Examiner on English 1 commented, therefore:

> and getting complete - completely bottom band marks is becoming - is becoming - er - very very rare so rare that I couldn't actually find an example of it in the limited time I had to search out these exemplars, urm, but that's the that's the normal case for preparing for [the] standardising meeting because the pool of resources is quite limited.

This direct comment demonstrates that he would usually consider it important to have a very low level answer, but the fact that an example was not easily found is in itself made telling; this also speaks to the *availability* heuristic, although if the lack of available example is because of the rarity value rather than another reason, it is a useful heuristic, and not a cause of bias.

It is possible to deduce from the practice of principal examiners that they also think it is important to select a range of essays which fall in the top band attainable, so that examiners can see that it is possible to get high marks without attaining perfection. This fact is made explicit in training, as for example in that for History 1, where the first Principal Examiner closes the plenary meeting with a comment on the third and final top band (level 5) answer of the morning with the following comment:

> you've got several scripts now that are in level 5, and you might go back and think well, this is a far far better level 5 than the William the Conqueror one. That doesn't mean that the William the Conqueror one is not level 5 - you see, it does display the necessary qualities to get into that level. The fact that PE2's two scripts demonstrate those qualities in spades is unimportant. They are obviously level 5. (PE1, H1)

The implicit context to this is that it is easy for examiners to under-reward scripts, and to think that it is harder to get into the top band than it is; all the other meetings encourage use of the 'full range' of marks. In another meeting a Senior Examiner comments that the mark schemes can be deceptive: 'because they always have to be positive they suggest an idea of perfection which is not attainable' (SE, English 2). One of the functions of the anchor scripts is to make concrete the abstract characterisations of the mark scheme, to provide a tangible comparison to the script at hand. One Principal Examiner used the word 'typify' to describe

the function of the sample material (English 1). Anchor scripts also demonstrate that the mark scheme does not have to be operationalised in just one way, so that the Team Leader can comment that a script is 'bad in a completely different way – just as you can get good marks in lots of different ways you can get bad marks in lots of different ways' (TL, H2). The selection of anchor scripts deliberately speaks to this range of routes to the same mark, according to the Principal Examiner of English 1, who introduces the plenary session by telling the examiners she has 'given a range of exemplar materials some of which show the pathways to achieving.'

More unexpectedly the transcripts reveal that principal examiners consider atypical or deceptive scripts to be good exemplar material, which would seem at first glance to go against the point and principle of an anchor script. In English 2 a script which the Principal Examiner characterises as 'a bit of a problem one' in that it is 'very short' turns out to be a very good script which despite the 'apparent brevity' has a great deal of content. One of the examiners comments, after they have agreed a high mark for it, that the script is 'deceptive isn't it – it would be great for a big team.' After some further discussion the senior examiners on the paper agree to set this answer aside for the summer examination session, when a much larger pool of candidates, and hence a much larger pool of examiners, will take part, in case another script with such good potential to be a sample does not occur. This reflects the fact that anchor scripts are dual purpose: not only are they designed to be comparative material and useful tools during the marking process, but also to provoke discussion and thought about what is actually valued by the mark scheme during the training meeting. Such a necessity exemplifies the difficulty of the task which examiners must undertake and reinforces the fact that the mark scheme cannot be a simple formula for decision-making.

All comparison during the meeting is with sample scripts, by the nature of the situation. In the cases of History 1 and 2 and English 1, they function as anchor scripts in the meeting just as they do during live marking, because the marks for them have been decided and set, and each successive script contributes to the frame of reference for what, for example, 'a level four looks like.' The scripts used in the plenary sessions at the beginning of those training meetings create a base line of reference from which the break-out group meetings can work later on. English 2 is a little different, because of the way in which the Principal Examiner ran the meeting, as discussed in 'Characterising the Training Meetings' above, in a more fluid and co-operative way. In that meeting comparison with the anchor scripts functioned much more in the way that comparison with other scripts does during marking, and as a result it is discussed below, in that section.

The use of the anchor scripts in decision-making varies from unit to unit, and they are barely used at all during History 1, neither during the plenary session nor during the break-out section, as far as can be established from what is said by the leaders of the session; that is, that once a script has been given a mark it is then for the most part left aside for the remainder of the meeting, rather than being used as an explicit touchstone when considering the subsequent scripts. There is an indication that despite this individual examiners are using the anchor scripts, although it appears that they are using them for a slightly different comparative purpose. Namely, examiners compare the mark they awarded with the mark awarded by the Principal Examiner and then adjust their later responses accordingly, in line with the gap they perceive between their intuitive judgement and the 'standard' as represented by the senior examiners. During the break-out session one History 1 examiner states that she was 'surprised' at the mark awarded to one of the common training scripts and so has been 'over rewarding anything approaching analysis' since then.

In contrast during the break-out session of the History 2 meeting comparison is constant between the scripts, although it is implicit in the number of comparative adjectives which the examiners employ, as opposed to obvious as a deliberate strategy. The Team Leader does direct her team at one point to compare the answer under consideration with the last one they marked, but it is mainly for interest's sake, although it also to point out that surface differences can detract from an underlying similarity: 'this one is all neatly written – [you] don't have the same problem as the last one with the glaring inaccuracies' (TL, H2).

What the meetings for these three modules (History 1, History 2 and English 1) do have in common is that when comparison is used, it is more in the nature of analysis, to elicit the characteristics of the different scripts, rather than to make summative judgements. The one explicit comparison made by the Principal Examiner on History 1 suggests that it would be interesting to compare the first paragraph of an essay with that of the previous one, because of its 'different approach':

> we are having an attempt at a multi-factored introduction here – not particularly expert perhaps but it is trying and when you compare that with the first answer this is, um er, a more interesting and perhaps more focused opening. (PE, H1)

The direct comparison is used to illustrate the qualitatively different, and better, approach which the second essay is taking. However, the second essay, despite the 'better' opening, ends up with a lower mark, which may suggest a reason why the overall comparison is not made. The more analytical approach is perhaps more useful to the examiners, as it enables

exemplification of what specific elements of the mark scheme look like, as opposed to what a '28' looks like; later scripts are unlikely to match an anchor in all features, but are quite likely to replicate one or two features.

During the live marking, use of the anchor scripts also varied between the participant examiners. The most striking use of anchor scripts is by Lupin (English 2), who in her earliest recording uses them almost systematically to confirm the decisions she has made, and to adjust those decisions if necessary. About half the time in the first hour she uses the sample scripts as an anchor for making serious manipulation of the marks awarded to a script, often locating an answer between two sample scripts before making a decision on it: 'it's not even getting to M it's more like – well it's obviously better than T.' (The letters are those given as references for the sample scripts at the meeting.) She also uses the sample scripts as a confirmatory measure so that later in the same session this passage arises:

> that's putting it above M. Not as far as L [9 second pause] but looking back over M... it's better than M – no it is better than M. What did M get? Thirty five. So no, that's not over marking it.

This is a more typical use of comparison in the data, to confirm or adjust decisions which have been made rather than as a tool for making the primary decision, and this function entirely dominates during Lupin's later sessions, although the use of anchor scripts diminishes too. It is notable too that Lupin uses comparison on the level of the mark for the individual Assessment Objective, rather than only on the level of the overall mark.

In contrast, Parrot does not reference the anchor scripts at all during her marking. This difference reflects both the difference between their respective training meetings (comparison is a constant feature of the English 2 training meeting, although it will be discussed below, for reasons already mentioned; it is much less used in History 1), and also potentially a difference in that English 2 was a completely new unit, which had never been examined before: the only frame of reference available was that supplied by the meeting and the sample scripts, whereas Parrot has the experience of previous examination series to draw on.

With other scripts

There is a suggestion from the History 2 transcript that the anchor scripts function in cohort to create an overall frame of reference in which an answer may be situated, rather than working as individual comparative touchstones, and that this frame is further fleshed out by

the live scripts once an individual begins marking. Thus one examiner adjusts her marking upwards during the training meeting because she has analysed her marking as 'too mean on the previous questions', adjusting her 'standard' on the basis of the anchor scripts. Her Team Leader, meanwhile, twice mentions the number 50 as a key point in the marking process, after which the frame of reference is in place and the marking process becomes 'more intuitive' and after which 'you've got it sussed.' This resonates both with Wiliam's 'construct-referenced' assessment (1998), Vaughan's suggestion that the papers in an examination session become 'one long discourse' (1990:121), and with the concept of migration of decision-making processes from System 1 to System 2, as with practice they become more automated. It also demonstrates that all scripts are marked in the context of a conceptual field of all other candidate's responses, and the relative as well as the absolute mark matters, so that the English Chair of Examiners can comment that 'the rank order is significant' and that 'you're not doing injustice to the other candidates if you give this one 10' (ChaE, E2).

The standardising meeting for English 2 is based around constant comparison, led by the Principal Examiner, and at the beginning of the meeting the comparison happens before the decision is made, at the level of individual Assessment Objectives, with the Principal Examiner asking his team 'AO2: is it better than the previous two?' At this early stage, in the first examination session of this paper, it seems possible that this is about building up the frame of reference in which to locate later answers, with comparison to more than one script as a beginning mechanism. Later in the same meeting the use of comparison occurs slightly later in the process, after a band has been decided, once there is potential material for comparison within a band, at which point comparison is used both for confirmation that an answer has been placed in the correct level, and then to make a firm decision about a mark. At the very end of the meeting the comparison is used as confirmation once all three Assessment Objectives have received a mark and the total has been calculated, for an overall comparison; now the team do not use the most recent example as a comparison, but the most relevant one, which answers the same question, and when the Chief Examiner comments that the two scripts are 'further apart than that' [i.e. than the marks would suggest], appropriate action is taken on one or both scripts, as necessary.

The reason that I have characterised the use of comparison during the English 2 meeting as being like that with 'other scripts' as opposed to with anchor scripts can be exemplified by the following exchange:

Principal Examiner: how does it compare with the others [in band 3]

	[Discussion establishes that this one is not as good as the previous one, which was awarded 8, i.e. the lowest mark of band 3.]
	Anyone like to hazard a…
Senior Examiner:	okay 9
Principal Examiner:	yes I've been sort of hovering between 8 and 9 because it says 'significant range of literary and linguistic features'
Chair of Examiners:	I think more and more we should go back and give the last one 9 and give this one 8

This standardising meeting is atypical in that the marks which the Principal Examiner has awarded to scripts before the meeting are not unchangeable. He makes changes at a number of points during the meeting, although it is not possible to tell exactly how much of the time, because he does not make explicit when he has in fact guided the team to his original mark. In this extract it can be seen that the marks awarded to the previous script are fluid, not set, and the going back to change the mark awarded to a previous essay in the light of new comparative information is typical behaviour in the use of comparison with other scripts, in that the previous essay is as likely to be changed in the comparison as is the new one. Despite the computerisation of the system, the examiners retain the ability to revisit the essays they have most recently marked, which means that they can adjust their marks in this way; this feature of the system suggests that it is an acknowledged technique in marking, and one which is not considered so damaging as to be worth preventing examiners from using it.

As with the ways in which the team in History 2 compared scripts constantly in a less formal way, the examiners in English 2 also naturally compare the scripts on a more generic level, with 'more promising', 'more sophisticated' and 'better' being mentioned as preliminaries to the serious discussion of marks.

Once again it is only Lupin, the marker on English 2, who uses comparison between scripts to any great extent in her marking. Occasionally she compares on a quite informal basis, for example 'that was a much better student', which is usually done to flag items which she finds of interest. In the decision-making process she uses comparison between two live

scripts much less frequently than she does between a live script and an anchor, and when she does it is not as a tool for primary decision-making, but as a confirmatory measure, post decision. On one occasion she totals the marks she is awarding and realises that although 'that's actually better than the previous one in many ways' she has given it the same total mark. She then embarks on a four minute deliberation on the relative values for each of the three Assessment Objectives for the two essays, which eventually ends with modification for the marks for both, before she finally compares the new overall essay total to the sample scripts.

With imagined scripts

The final touchstone is a somewhat unlikely one, although its existence has been argued elsewhere by Alistair Pollitt (2010) who alleges that in the absence of physical scripts to compare, examiners will resort to ideas of what a given level 'looks like'. It is also suggested by the 'prototypes' of Crisp's (2010a) model, the 'mental models of likely typical responses.' There is certainly much comparison throughout the data with hypothetical scripts and the change in marks that would result if a script looked a little different; examiners also constantly refer to what 'a better candidate' would do, or what an ideal answer would look like.

The general principle on which contemporary examinations are marked is that of positive marking; i.e. candidates are rewarded for what is present, rather than being penalised for what is wrong or missing. This is generally acknowledged, and sometimes examiners are explicitly reminded, particularly when it appears that someone might be moving away from that, so that the following reminder can be issued to a team of examiners:

> We're marking what they've done. We're not doing that thing where we diagnostically mark… that's not fair. (TL, H2)

In other words, examiners should not impose their own idea of what content should be in the answer. The mark scheme for all the units in this study includes a section called 'indicative content' for each question, but it is emphasised in all the training meetings that it is 'indicative' only; i.e. it is neither an exclusive list of acceptable content, nor is it necessary to mention all the points mentioned to get a top grade; its widespread neglect was noted in the previous chapter. Despite this the Team Leader on History 1 says to his team, after discussing two top band answers:

> I think the weakness of both of the ones that we've looked at so far was something that you know, really top candidates will have actually seen in that question which wasn't addressed by either of them. Would you sort of think that there's anything that both of them missed out on that you might be expecting from a really strong candidate?

His team pick up on the factor to which he refers, and begin to worry that an ideal answer would address something that these candidates are not. In response he backtracks a little and says:

> obviously I think we need to keep in touch on that always though, because it may well be that virtually every candidate, er, interprets the question in that way and er, er, we don't want to debar people from getting a really high mark, if everyone's interpreting it that way. I think we need to keep in touch, on the way people actually answer that question.

The comparison with the idealised answer is subordinated to comparison with potential actual answers (incidentally also showing that the relative merits of answers also matters on History 1, despite the lack of explicit comparison during either training or marking). This comparison is still, however, with imaginary answers. Elsewhere the ideal is less specific, so that a script may be described as 'not as well developed as it might have been' (PE2, History 1) or an examiner can comment 'better scripts would have said that it wasn't archaic at the time' (Senior Examiner, English 2). Similarly the Team Leader in History 2 remarks of one script 'they could have done rather more with the sources', with this lack being cited as the reason for its not receiving full marks.

Imaginary answers are referenced elsewhere in the data, and not only idealised ones. More common is the question from examiners as to whether, if one given feature was stronger, it would be enough to justify a better mark. An examiner during the plenary session of History 1 asks just this, in reference to the ever difficult 'Quality of Written Communication' which is not specifically delineated in the mark scheme, although it is supposed to be allocated two or three marks in the overall total. The exchange goes as follows:

> **Examiner:** [If the] quality of written communication been better would that have got erm, higher up in level 5? Possibly middle or top...

Principal Examiner 1: Ah. Thanks for that question Seamus. Certainly not to the top, certainly not, because quality of written communication is a marginal matter for you to consider. If it had been better or more expert I think it would have been worth... 26... but I think to go down to go even higher than that would have placed the er script perhaps in the wrong position.

The use of the hypothetical comparison enables the Principal Examiner to explicate the relative importance of one characteristic of the script, in this case, being worth only a single mark. How much better it would need to be is, however, left unclear, which is the difficulty with the hypothetical comparison: defining what is meant by 'better' or 'more expert' is very problematic, and indeed a main function of anchor scripts is to exemplify what these abstract comparatives in the mark scheme in fact mean.

Interestingly, the Team Leader on History 2 is unwilling to give answers to such questions featuring hypothetical scripts, refusing explicitly on two separate matters to issue a definitive ruling, warning them that she's 'very nervous about categorical [statements about what is required to reach a given level]' and later 'I'm not going to give you a straight answer on the interpretation of sources' in response to a set of 'if' questions from her team. She is keen not to give hard and fast rules which may lead to error if followed unthinkingly.

There is another hypothetical comparison which is made in every single unit in one form or another, expressed in its most classic form as 'we have to ask ourselves, could a seventeen year old do any better than this?' (PE2, History 1), which requires examiners to limit their imagined ideals in terms of the candidates' demographics, and in other places, in terms of the time available.

The anticipated shape of the field of answers is also considered, again showing that each answer is considered relative to the entire field; the question 'are we going to see ones which do have more on the language?' is answered with 'I don't think we will' (Senior Examiner and Principal Examiner, English 2); although the Principal Examiner has a slightly greater idea of what the field looks like, it is still not enough to make this much more than speculation.

Where the anchor scripts do not completely exemplify a given phrase in the mark scheme, it is natural that examiners must have some mental concept of what it looks like in practice, although this is more frequently seen in the data as they wonder whether a given script does in fact represent, for example 'some awareness'. This occurs in a number of places in the 'think aloud' recordings from the live marking. Interestingly while Lupin compared extensively to anchor scripts and other scripts during her marking process, and Parrot did not, this pattern is completely reversed when it comes to imagined or ideal scripts. While Lupin never makes comparison to anything other than a concrete script, Parrot repeatedly explains her marks in terms of what has not been included, which a better script would: 'he didn't mention the Duma'.

Primary decision-making/ confirmatory

It is clear that while comparison can be used in an analytical way, to explicate the different aspects of the script, in terms of decision-making comparison occurs in two distinct locations, at the point of the primary decision, and secondly after a primary decision has been made, in a confirmatory function. Some types of comparison, such as that between a candidate's answers only occurs post-hoc; the use of other comparison, particularly with anchor scripts and with other scripts occurs at both points. Returning to the discussion of the potential for a halo effect with comparison between the two scripts of a candidate's answer, it is possible that such comparison only occurs at the confirmatory stage because of the awareness of the possibility of consequent bias if it occurs at the primary decision-making stage. Alternatively the fact that it only occurs at this stage may prevent bias (which is arguable, given the tendency of Parrot to alter the original mark subsequently). Or yet again, it may occur unconsciously at the point of the primary decision, but is only mentioned explicitly at the point of confirmation.

There is one type of comparison which can only occur at the confirmatory stage, which is the comparison between the marks awarded for each Assessment Objective and the overall mark for a question, against the 'feel' of the whole answer. In particular English 2, in which the examiners have to cope with three different Assessment Objectives, which are differently weighted, has the problem that while examiners have to 'stick to the AOs', they also need to consider the 'total mark' (Principal Examiner). The analytical marking makes this type of confirmatory check difficult, because one cannot simply add a mark to the total to compensate for it seeming under-marked; the mark must be added to a specific Assessment Objective. There is a similar comparison between the overall mark and the 'feel'

of what the answer is worth in History 2, but it is more at the level of the pass/fail boundary, and more often in that unit, despite the sympathy expressed, the confirmatory check does not result in a change in the marks awarded, although it is used to warn examiners to give the benefit of the doubt when debating between two marks in their own live marking. Thus what is a confirmatory check in a training meeting is used to teach a lesson about the primary decision-making.

Both the English 2 training meeting and Lupin's subsequent marking of that paper show a movement from an initial use of comparison as a primary decision-making tool to the use of comparison to adjust or confirm a decision which has been made, later in the session. It is interesting that her live marking represents the same processes in microcosm which the whole team went through during the training meeting. Initially the other scripts are used to establish the band in which an answer belongs, before movement within the band is decided by closer examination of the features of the script and the mark scheme. (This reflects the mandated practice of the History 1 paper, for which examiners are instructed to locate a band and then move up or down from the middle mark of that band. It also reflects the *anchoring and adjustment* heuristic in action.) The middle stage uses the comparison after a band has been decided but before a mark: by this stage enough scripts are available that a frame of reference has already been established (potentially implying the unthinking comparison which is examined below) and it is possible to compare relative marks which are just one or two different, instead of having to compare on a larger scale. Finally the comparison moves from the primary decision-making process to the confirmatory one, where a mark decision is made, and then comparison is used to adjust the mark as required.

Rule-based comparison / unthinking

Some comparison is mandated by the senior examiners, and this is mostly that which is described above, both in terms of the primary decision and the confirmatory check. In some cases it is explicitly suggested as an approved behaviour, as in the heuristic offered in History 1 that both answers tend to operate at the same level. In others it is simply modelled behaviour, as exemplified by the English 2 standardising meeting. Lupin's systematic use of the sample scripts in her earlier 'think aloud' recordings are suggestive of a rule-based approach to the technique; the fact that comparison to the anchor scripts becomes far less common as she proceeds through the marking period indicates that she adheres to this practice only as long as she feels she needs it. All three examiners who took part in the 'think aloud' recordings used comparison of some description, whether with other scripts, sample

scripts or imaginary scripts, or indeed, with the mark scheme, as described in Chapter 6; it is particularly a technique which they use when having difficulty making a decision, especially in the case of the mark scheme. This suggests a conscious resort to comparison, and an awareness of its usefulness.

Implicitly, I have followed the model established by Baird (2000) in this consideration of comparison, which suggests that decisions on scripts are made in the context of a mental framework that represents the range of attainment possible on a paper, and includes conceptions of what a given grade, or band, looks like. The movement in English 2, for example, given that it is a new unit, in which examiners cannot have any prior experience, fits nicely with the idea of the different scripts creating a comparative framework which they build on with each new script, until they are able to compare new scripts to their mental conceptions without explicitly or consciously doing so. This is supported by the History 2 Team Leader's assertion that after 50 scripts on a paper the marking becomes easier, which could be because the new question has been slotted into the mental map. (Alternatively it could also be because of a much more prosaic System 1 to System 2 migration, in which examiners have simply 'internalised' the mark scheme and committed it to long term memory, making the process much less effortful.) It also fits with the lack of direct comparison between scripts, whether sample scripts or otherwise, on History 1 and English 1, which are established units, being marked by experienced examiners, in which the comparative framework is already fleshed out, and the comparison can be unthinking rather than explicit. This is supported by Lupin's diminishing use of the sample scripts in her 'think aloud' recordings.

There is another kind of unthinking comparison, however, which is that which does not relate specifically to decision-making, instead making comparisons of the following sort:

> More sophisticated (Examiner, English 2)
>
> A better candidate (Caspar, History 1)
>
> No worse on this (Chair of Examiners, English 2)

Examiners make unthinking and generic comparisons throughout the training meetings and the 'think aloud' recordings. The word 'again' recurs frequently, and implies that common features are being identified, which requires that unthinking comparison is occurring. This in itself is evidence that comparison is a natural feature of human judgement, and a natural response to being presented with a succession of different examples of anything. The rule-based and explicit comparison may have harnessed that process within the standardised, mechanised process, but it occurs beneath and beyond that in every one of the units.

Conclusions

Comparison is clearly one of the main decision-making behaviours in which examiners engage. It is an automatic process which happens at all points in the consideration of scripts, but which may be harnessed within the formal remit of the examination marking system in terms of the anchor scripts. It is also used in less officially certified ways, as with the comparison between different candidates' scripts.

There is a clear recognition that comparison occurs, and the comparison within candidates' scripts, between Assessment Objectives or between essays, is governed by the warnings which senior examiners give to their teams. Comparison with idealised and imaginary scripts or candidates occurs frequently and often unconsciously; some senior examiners warned against it, while others, apparently inadvertently, suggested its use. It is possible that creating more awareness of this type of comparison might prevent its misuse. Making explicit the comparison between Assessment Objectives may mitigate a potential halo effect, and also any bias due to the *anchoring and adjustment* heuristic. Comparisons between the total awarded via an analytical mark scheme and the perceived holistic value of the script may also help examiners to reconcile the two sides of that debate, and their own desire for what they see as more valid assessment with the Awarding Body's need for reliable marking.

Comparison is particularly prevalent at two points in the marking process: at the point of decision-making, particularly early on in the marking period, or when the judgement is a difficult one to make; and as a post-hoc confirmatory check, to ensure that the correct mark has been awarded. It is suggested as a way of checking that the right decision has been made by a number of senior examiners, with a variety of comparative material suggested. This is the use of the mark scheme which was highlighted in the previous chapter.

Unconscious comparison to some extent forms the basis of the cognitive heuristics which will form the basis of the next chapter. It is also the key mechanism by which construct-referenced assessment must take place; the concept of a mental framework into which a script slots relies on the instinctive comparison of that script with all the other reference material making up that framework. The evidence presented in this chapter suggests that examiners certainly engage in instinctive comparison, with a wide range of comparative material.

Chapter 8: Heuristics – cognitive shortcuts for human judgements

"So convenient a thing it is to be a reasonable creature, since it enables one to find or make a reason for everything one has a mind to do." Benjamin Franklin

Heuristics have become widely accepted by social psychologists as a mechanism by which rational judgements may be made quickly, by circumventing the rule-based, careful processes of conscious decision-making. As outlined in Chapter 1, the four formal heuristics which will be considered in relation to the data are *representativeness*, *availability*, *anchoring and adjustment* (Kahneman *et al.*, 1982) and *affect* (Slovic *et al.*, 2002). The three posited by Kahneman *et al*. (1982) are examined first. There is some evidence for the application of each of these by different examiners at different times, and there is also clear concern for the ways in which they may lead to bias, in the instructions which principal examiners give as a pre-emptive counter-measure.

The cognitive heuristics are unconscious, but the examiners' attention is drawn to them where they may mislead judgement. It seems likely that the experience of Awarding Bodies, or of individual principal examiners, has led to adjustments in the system, particularly in the selection of sample scripts, which are designed to mitigate the potential bias caused by heuristics, though not necessarily consciously. A typical example is that of the emphasis on very high achieving scripts in the samples provided to examiners, clearly a result of the fear of systematic under-rewarding of scripts which are not 'wow' scripts (as the Team Leader on History 2 expressed it) but which are top-level nonetheless. This is combating potential bias caused by the availability and representativeness heuristics, but is likely to be a response to a specific concrete situation as opposed to an abstract theoretical problem.

A number of informal heuristics, specific to the situation, perhaps more properly characterised as 'rules of thumb', given to or used by examiners, have also been identified. The evidence for their existence and their relation to the formal cognitive heuristics will be outlined. Informal heuristics are often accompanied by hedging or other warning statements to prevent their over-application.

Finally I will examine the *affect* heuristic; examiners demonstrate a variety of affective and emotive reactions to the scripts and their projections of the candidates behind

the script, but accompany these with hedging behaviours and statements designed to show that they will not let these affective reactions skew their judgements.

The cognitive heuristics

Representativeness

To recap, the *representativeness* heuristic proposes that quick, accurate decisions can be made on a smaller number of characteristics, or even a single characteristic, which are seen as being representative of the category into which a script (or other item) is being put. The phrases 'looks like' or 'feels like' are characteristic of the application of this heuristic, and it is particularly in evidence in relation to the top level scripts. Other phrases also suggest that examiners characterise the 'feel' of an essay in a variety of ways which enables them to categorise the scripts as an aid to putting them in the correct levels. The first Principal Examiner on History 1, for example, describes a script's approach as 'a general studies way'; the examiners in the standardising meeting for English 2 were particularly concerned with the 'checklist approach' towards writing the essays. Both of these characteristics are representative of lower levels of achievement.

The length of quotations is clearly representative for the English examiners; Lupin frequently remarks on the presence of over-long quotations without accompanying explication by the candidate, and during the English 2 meeting she commented 'I was a little uneasy about the length of the quotations.' Quotations were an easy source of representative characteristics for the English examiners, who also regarded the presence of 'nicely embedded' quotations as representative of a higher order of skill.

Length is clearly considered to be a representative characteristic, in particular shortness is representative of lower grades. The following exchange in the English 2 meeting exemplifies this:

Principal Examiner:	can we look at J? I found this a bit of a problem one - it's a very short answer
Examiner:	Oh golly [*they read*]
Examiner:	Good stuff though. Hard to take exception to any of it. Hardly a word wasted. I can't believe that wasn't planned. They probably spent more time thinking about it than writing

Principal Examiner:	But it's short isn't it
Examiner *:	It is but it's small writing
Examiner:	It's literally getting twice the number of words to the line
Senior Examiner:	It's discriminating, no doubt, not wasting a word

Although the Principal Examiner is focused on the shortness of the answer, the examiners participating in the meeting defend it not only on the basis of the content but also on the grounds that it is not as short as it appears, because of the handwriting. They continue to delineate the essay's positive points after this exchange, until at the decision the Principal Examiner admits 'I was thrown by the apparent brevity' (which is in accord with Crisp's (2010b) finding that when the length of a response does not correspond to an examiner's expectations, then that also becomes a factor in their decision-making)**.** It is this essay which the team decides is 'deceptive' and put away for the summer standardising meeting, as discussed in chapter 4 above. Despite its apparent length, the essay is placed in the top band. It was clear that the Principal Examiner had intended the essay to be a sample of a poor essay, and his emphasis on the length indicates that this was the characteristic which had led him to this conclusion; it is also clear that the principle that shortness is indicative of low quality is not overturned here, as the examiners establish that the essay is longer than it looks. A similar case emerges during Lupin's 'think aloud' data; she notes 'it's quite short' 'very tiny writing' and subsequently goes on to assess how much detail is included to decide whether it is apparently or actually short. Parrot similarly comments on an essay which has 'not enough detail so it's low. And in addition it's a very short essay with a lack of information', a summation which places it in the middle band.

What is interesting to note is that representative characteristics are not usually representative of a specific band, but rather of a higher, lower or middle ranking essay. The only characteristic which is repeatedly connected to a specific band is the word 'adult', as in 'this is written by an adult' (metaphoric in use, since no characteristics of the candidate are available to examiners) or 'the writing is completely adult'. This is a term of approval which is linked to essays in the top-band, and particularly those given full marks. Parrot is the most notable exponent of it, but it was also used in two of the standardising meetings. In general, however, representative characteristics do not appear to be tied to levelling decisions, which makes them both of less use to examiners and less likely to lead to bias. It may be related to

the fact that although there is intended to be a qualitative difference between bands (considered further below), examiners in fact see it in terms of gradations rather than categories. No one representative characteristic was seen to drive any single decision, although one was particularly influential in adjusting the decision: Quality of Written Communication (QWC).

Quality of Written Communication

Although the quality of the writing is formally a 'marginal matter' (Principal Examiner 1, History 1), it is an aspect of the scripts which is frequently noted by examiners as they discuss the essays during the standardising meeting, and during their 'think aloud' marking sessions. It is only on English 2 that the quality of writing becomes a formally assessed matter, as one of the Assessment Objectives includes the requirement for 'accurate, coherent written expression.' It is an Assessment Objective with a low weighting, however, which is responsible for a very small proportion of the marks available for that paper.

The examiners on other papers do, however, use QWC as an indicator of the quality of the essay. The same Principal Examiner who dismisses it as marginal, later includes QWC in his summation of scripts, even using it to balance something negative:

> 'Now you may think that's not a very good conclusion, well, the candidate is trying to weigh up the factors that he or she has already considered and we do have here an answer that makes five points, has secure quality of communication and has a conclusion.'

He is attempting to steer them towards a higher mark, and this script is one which appears in the top band. Elsewhere comments that an essay is 'deeply badly expressed' or 'flowed nicely' (both History 2 examiners) are both seen as justificatory support for the decisions which are made.

Lupin, an examiner on English 2, frequently comments on the 'style' of the essays she marks; however in her decision-making for the Assessment Objective which mentions 'written expression' she focuses exclusively on the other elements in the AO description, particularly the range of literary and linguistic terminology which has been used. The 'style' comments form instead a general impression of the quality of the essay early on; for an English examiner this may be closely related to, and aligned with, the skills which are actually supposed to form the basis of the assessment. Parrot's use of QWC is not as a representative characteristic, however, but instead as a means to adjust the mark once she has made an anchor point; she decides on an area of two marks and then says 'I'll give it 15

– it's quite well written.' (Interestingly, this contradicts her earlier comment on the essay that it was 'not very well written', suggesting that QWC is not a major concern for her. Her marking is considered further under 'Informal Heuristics' below.)

Essays which are excellent, and worthy of full, or almost full marks, however, are often identified first by their writing style (related to the discussion of 'adult' above), as an easy marker. 'Fluid' and 'sophisticated' style are both seen as representative of the top level. The best practitioners of QWC are also likely to have structured their essay, argument and content, to the best effect, so that it becomes a proxy for the other skills which are formally examined on the mark scheme. Examiners are warned against candidates who appear to be demonstrating those skills through their use of signposting using discourse markers, for example, but who are not, so they should be careful to establish when the QWC is commensurate with the other skills and when it is not.

For the standardising meeting on English 1, the quality of written communication also becomes an integral part of the commentary, particularly where it may be deceptive ('a bit wordy, but it's talking about subject specifics'), which suggests that there is an acknowledgement on the part of the Principal Examiner that these characteristics may be used to make judgments of scripts. Similarly during the plenary training the Principal Examiner for History 2 warns his examiners against relying on the language of the writer to make a judgement, commenting that while an essay is 'perhaps not the most fluid of scripts you will read – it's quite clunky', it is nevertheless situated in the top band.

Quality of Written Communication is a characteristic of scripts which is relatively easy to see and to judge, as an automatic process, particularly for the English examiners. It seems likely that for some it becomes a representative characteristic, and as such principal examiners feel the need to warn against its potential for causing bias if wrongly applied. It has been noted, however, that it is not a characteristic which drives decisions on its own, although it is influential. It seems rather, to be brought in as part of the adjustment process connected to *anchoring and adjustment*, which will be discussed below. In that way, it is a 'marginal matter', and the comment from the Team Leader on History 2 that 'technically there are two and a half marks somewhere' becomes more realistic. It is worth noting that the data in this study did not support Crisp's (2010b) finding that language became a more manifest criterion when there was some difficulty with the way in which a candidate had expressed themselves, with corresponding problems with the interpretation of the answer by the examiner: it was always a manifest criterion.

Availability

The evidence for the *availability* heuristic is particularly difficult to discriminate in the data. More than most, its presence is unconscious and unlikely to be expressed. That it is present, and applicable, however, may be deduced from the behaviour and speech of the Principal Examiners, in particular. Principal examiners repeatedly demonstrate their concern that the *availability* heuristic may lead to bias, in their provision of multiple examples of top-level scripts. A Principal Examiner on History 1, for example, makes a point of this at the end of the plenary:

> you've got several scripts now that are in level 5, and you might go back and think well, this is a far far better level 5 than the William the Conqueror one. That doesn't mean that the William the Conqueror one is *not* level 5. You see, it does display the necessary qualities to get into that level. The fact that PE2's two scripts demonstrate those qualities in spades is unimportant (PE1, H1).

His concern to show that all three are top band scripts suggests the intent to provide enough different examples at the highest level that examiners do not under-mark for a lack of available examples – or indeed a misconception of what is representative of the top band. This concern is mirrored in the sample scripts produced for the other history module and also for English 1. English 2 is an exception, in that it is being marked only by senior markers, and is a new module. The sample scripts selected for it are designed to cover the widest possible range of achievement, and of questions, presumably because it is more important to ensure the availability of a range of examples before beginning to be concerned about potential bias for lack of scripts at the higher end. They do demonstrate an awareness of the need to use the 'full range' of marks, but at a first examination session the Principal Examiner cannot be sure that encouraging markers to award the top band is an appropriate move.

Similarly the setting aside of a deceptively short script, considered under *representativeness* above, also suggests a concern in the selection of sample scripts to provide examples of unusual but still possible cases, to prevent available examples producing bias. In fact, one Principal Examiner apologises for the lack of samples gaining the lowest band marks, because he was unable to identify any before the meeting: this lack of availability for the Principal Examiner, let alone the individual examiners, would seem to suggest that the heuristic will steer examiners towards the middle and higher bands. However, if there are genuinely too few examples for one to be easily available at decision-

making, that in itself is indicative that any given script would be unlikely to be correctly located at a below-pass mark: it is not a cause of bias, but an appropriately working heuristic.

Another moment which suggested the application of the *availability* heuristic was the moment of decision-making as seen in the 'think aloud' protocols, immediately after a mark had been initially suggested. Examiners would think through the potential comparative material to find the most useful touchstone; the sample scripts, together with the scripts which had just been marked, provided the most *available* examples. The systematic use of comparative material, as discussed in Chapter 7, to some extent negates the *availability* heuristic, as it is not mental experience which is required. However, the comparison becomes very quick at times, far quicker than could be done by an actual comparison of the two scripts, so that Lupin, for example, says 'okay, three and five and eleven. Is nineteen. Just checking, that puts it just about T, yes that's okay.' The sample scripts to some extent must dictate what the cognitively available examples are, which puts still more of a burden upon their selection. It should, however, prevent to some extent the other scripts which an examiner is assigned having too great an effect on the marking of any given script, something which was more necessary before the computerisation of the marking process created random assignment of essays, instead of whole schools to a single marker.

In Parrot's first 'think aloud' session, all but one of the essays was judged to be a band 3 answer, and the majority of them as mid-band 3 at that. Yet the sample scripts for her module were heavily biased towards the top end, with the plenary scripts providing three top level scripts out of six. We can presume that her marking is at an acceptable level, given the nature of the sample, so the majority of band 3 judgements is presumably accurate, particularly given that it did not occur in her other recordings. *Availability* is not the only heuristic at work, perhaps; sample scripts can mitigate its effect in causing bias, particularly in reducing the numbers of top level scripts which go unrewarded, but are not required to cover the entire range of marks in the same depth. Alternatively, as experienced examiners, the participants in this study would have sufficient available examples from their own past experience to be able to judge accurately.

Anchoring and adjustment

There is plenty of evidence for the use of *anchoring and adjustment* by examiners; it is the typical way examiners work, consciously choosing a central mark and then adjusting up or down in light of the mark scheme, sample essays or other live scripts, as discussed in the

previous chapter. In one of the modules the Principal Examiner explicitly tells the examiners to use this process, but it is not markedly more common in the data from that module than from others:

> once you've decided on the level suited to the answer can you please begin to look from the *middle* mark point in the level. The generic mark schemes states straightforwardly that the descriptor of the level attainment is the *middle two* marks, and then look to move up and down from the middle mark of the level not from the bottom.' (Principal Examiner, History 1)

Its use is supported by the structure of the training meetings in all subjects, which begin by asking examiners to find the correct 'band' as an anchor, before moving to choosing specific marks for later scripts. All examiners followed this procedure to some extent in the 'think aloud' protocols, particularly at the beginning of their marking period. After a while most would move straight to selecting a mark without confirming the band first; Parrot was the most notable practitioner of this. She was also the examiner most likely to change her decision on a band after having made it, usually on the band 3 and band 4 borderline, and almost always downwards. The examiners on the other history module (H2) had been instructed not to go between bands once they had decided on one, and to reserve the bottom mark in a band for 'dithering'; it seems likely, given the explicit instructions to the examiners on History 1 quoted above, that this would apply to Parrot's paper too. If the bands are qualitatively different, then a paper should either be one or the other, not debatably either. However, one of the things which is clear throughout the data and the discussion of it in the preceding chapters, is that examiners do not necessarily find the bands to be qualitatively different, nor the judgements to be correspondingly easy.

It was not the only way in which the *anchoring and adjustment* heuristic was in evidence, however. During the 'think aloud' protocols even after a mark within a band had been selected, the examiners would go through a substantial period of checking their decision, usually by the use of comparison, as described in Chapter 7. This was the point at which the mark scheme was sometimes brought back into the process. Lupin, for example:

> Critical understanding, or fully and critically, there would have to be a lot more to be fully and critically – what about, there's just not enough there is there? It is analytical, there is some critical understanding, it's more than limited, it's 6 or 7 again [*pause*]. I'm not sure that's a 7 you know, I think that's a 6, look at the spelling

> of marriage [*pause*] prejudice [*pause*] oh I don't know, you know there's a good bit of context there isn't there. It's not fully and critically, I think it's probably a 7. I wonder if I'm over marking it [i.e. over-rewarding it].

Having decided on a mark on a single Assessment Objective at this length, when she had assigned a mark for all three AOs, Lupin again went back to adjust the overall total in the light of comparison with the sample scripts. Adjustment was also occasionally retrospective, in that an examiner would revisit a script which had already been awarded a mark, in the light of the mark decision for the second essay in that script, or for a subsequent candidate's essay. As noted above, the Quality of Written Communication was frequently cited at the point of adjustment, although more often in the 'think aloud' protocols than in the standardising meetings. Caspar, for example balanced the Quality of Written Communication against other factors: 'it gets 16 since it's quite well written but it's missing half the question.' The senior examiner seem to approve its use as a driver of adjustment rather than of anchoring; the Team Leader in History 2 tells her team 'if you're dithering and it's appallingly spelled go a little down, but it shouldn't shift it a level.' It is clear that in places the Quality of Written Communication does provide the initial anchor, but essentially only when an essay is outstanding already.

Some research into anchoring and adjustment has suggested that priming an individual with a given number can startlingly affect the decision which they subsequently make, usually within estimating tasks (Jacowitz & Kahneman, 1995); individuals do not sufficiently adjust their estimation after fixing on an anchor. It would be possible to argue that Parrot's attachment to band 3, as described above, was an example of anchoring bias, having selected a mid-level anchor (which might be seen as appropriate, given the instructions to begin in the middle of any given band), she does not move far from it. The counter to this is that she was an experienced examiner, with years of acceptable judgements behind her, and she successfully completed the marking session which was the basis for this study. The practice on History 1 of not only using a band as an anchor, but also the middle mark within that band, means that to some extent the anchor cannot be inappropriate further than a range of approximately two marks, which is very small in the context (and most trainers tell their examiners that a disagreement of one or two marks is not going to 'bother' them, as the Team Leader on History 2 put it). It is presumably for this reason that an institutional version of the anchoring and adjustment process has been mandated for examining judgements.

Unofficial heuristics

A number of unofficial heuristics are identifiable, which senior examiners provide to examiners as 'shortcuts' through the official process for marking judgements, sometimes accompanied by specific numeric rules about marks. The unofficial heuristics are considered with the framework provided by the heuristics of Kahneman *et al*. (1982; 2002) and where appropriate, assigned to categories of heuristic within that framework. These heuristics are not cognitive heuristics, in that they are simple 'rules of thumb' which are easy for examiners to use, but which are very definitely conscious in their application. However, when combined with the theory of System 1 and System 2 judgements, and the migration of certain decision-making procedures, it is possible that these informal heuristics do in fact become less conscious in their application, which is the moment when they could begin to lead to bias, therefore becoming far less useful.

The most obvious shibboleths provided for examiners were in the history papers. The 'stated factor' of History 1 and the sources of History 2 were both impressed upon examiners as being the most important focuses of attention on which to base a decision, but for both 'analysis' was also given as an informal heuristic. Thus, on History 1,' as long as they're attempting any analysis at all, that's the crucial point to get to level 3 isn't it – if there's any analysis attempted it's going to be a level 3 answer'. Other firm rules were given, such as 'if there is no own knowledge level 1 is where it's going to fit' for History 2. These rules are related to, but not made entirely explicit by, the mark scheme. They provided a simplification and a hard and fast rule for the examiners which made the process of deciding on a band, if not a mark within it, easier. A rule specifically relating to marks was given to examiners on History 1: 'an answer that does not address the given factor cannot go above 18 marks' (a mark in the middle of a band) according to the second Principal Examiner on that paper. This very specific rule was extremely unusual, as most were either slightly vague, or accompanied by hedging statements, as discussed below.

It was more difficult for English examiners to receive hard and fast rules about what must be present for any given level. Although the same concern with 'analysis' existed, which has been considered elsewhere in this thesis (Chapter 5), the main focus of the questions tended to be expressed in the appropriate Assessment Objectives, so that 'context' for example, was specifically assessed, rather than being a clue to use to locate the essay in a more generic description. There was a general agreement on English 1 that essays must deal with both texts strongly to be able to reach the highest level; both English

meetings commented that moving between the texts throughout the essay was not a characteristic which could be used to judge an essay automatically, as 'they know they have to' (English 2) or '*we do not have to* see the comparison' (English 1).

Some examiners were clearly using other rules of thumb on which to make their decision. For Parrot her main concern was the amount of 'information' which an answer contained; every answer is weighed on that basis, to the extent that she could be categorised as Vaughan's 'single-focus rater' (1991:118). A typical comment was 'I'm going to give it level 3 because there's a lot of information and I think that's rather good.' There was no systematic quantification of the amount of information required for each level, though: another script was 'verging on low level 4 because basically it's got some good information but not enough.' The word 'detail' also appeared in this manner. Sometimes she linked it to the given factor which was a more approved potential heuristic, but not at all times, and rarely by that term. So, for example, the essay which has a 'lot of information' above is balanced by the fact that there's 'not enough about the Duma' (the 'stated factor' for that question). Though her concern with information was unusual and striking, it is not outside the bounds of the concerns of the subject; the first Principal Examiner for History 1 praised an essay that was 'covering an amazing amount of information in a relatively succinct form'. History does, after all, require factual knowledge with which to construct the arguments and evaluations which are the higher levels of skill in the subject.

Most of the unofficial heuristics which are given could fall into the category of representativeness, in that they might be considered to be the characteristics which are the most predictive of the correct level. Certainly the informal heuristics given by senior examiners fall into this category, in that they note certain features which are representative of certain levels. The 'information' heuristic used by Parrot is more a reduction of the judgement process to a single factor, which while it simplifies, still requires a qualitative judgement to be made. However, all the informal heuristics have one thing in common: initially they are conscious rules, not unconscious cognitive processes. Parrot's information heuristic is a possible exception, in that she may not know that she is using it, or at least so heavily.

Hedging

The informal heuristics which senior markers give to markers are not usually given unconditionally. They are usually accompanied by hedging statements from the senior examiners, to prevent too heavy a reliance on such rules. The Team Leader on History 2, for

example, tells her team 'If there was no awareness of where it might have come from I would be hard-pressed to give it a level 4. I'm very nervous about categorical [statements] but *provenance*.' The Team Leader on History 1 is similarly reticent about hard and fast rules during the discussion relating to the expected content of high level responses; having shown his team that 'really top level candidates' should be considering a certain aspect of the question, he hedges:

> obviously I think we need to keep in touch on that always though, because it may well be that virtually every candidate er interprets the question in that way and er, er we don't want to debar people from getting a really high mark, if everyone's interpreting it that way. I think we need to keep in touch on the way people actually answer that question.

Another such a heuristic was given by the first Principal Examiner on History 1, who does not present a definite rule, but a suggestion of a situation in which examiners should reconsider. He suggested that:

> if you've awarded marks for both answers which are *substantially* different from each other it is worth reflecting on your decisions before moving on. In the summer in general it became obvious that most candidates tended to approach both answers at the same level of answer.

The instruction provides gentle guidance but leaves the final decision to the examiner's own judgement, so that they are merely reconsidering decisions, rather than changing them in an automatic way to match the rule. Related to these informal heuristics is the identification by principal examiners of signals within the writing of candidates which indicate the presence or otherwise of specific qualities. The plan, or the use of discourse markers, could form a shortcut for markers. Similarly these signals were often accompanied by warnings that they could be false, or a level of uncertainty if they actually were accompanied by the structure they indicated.

I would suggest that these hedging statements are designed to ensure that unofficial heuristics do not lead to systematic bias. They create an air of 'productive uncertainty' (explored further in Chapter 10) for the examiners, who have some additional guidelines to make it easier for them to make judgements, but not to such an extent that they become over-confident. It is also possible, given that the strongest reluctance comes from team leaders, rather than principal examiners, that they are reluctant to produce

statements to which they may be held if examiners misuse them later, in that they hold a role in which they are both responsible for their team, and supervising their judgements.

The affect heuristic

There is considerable evidence for the existence of affective and emotional responses in the examiners. Many scripts prompted emotional reactions from examiners, often of frustration on account of poor work or of excitement when a better script appeared, or even when an unexpected improvement was demonstrated mid-essay:

> then luckily we get onto children - education.... okay... and here it's much much better, thank goodness, bringing it up again to middle, well, level 3 (Caspar).

Similar pleasure was evinced when a student had produced a plan, which most examiners were pleased to see, and which they regarded as a sign that a good essay was likely to follow. This could easily lead to disappointment, however, if the essay failed to live up to the plan, as Lupin occasionally found: 'there's a *lot* more in that plan isn't there [*pause*] that's a pity, a real pity.'

There was limited evidence of Vaughan's (1990) 'laughing rater' but a counterpart emerged much more prominently: the 'tutting rater'. Throughout the 'think aloud' recordings all three examiners had a tendency to use non-verbal utterances to express disapproval, but one in particular (Lupin) made tutting sounds extensively through scripts, sometimes accompanied by despairing laughter. At one point she packs in two heavy sighs and three tuts into one minute of reading. Many of the examiners during the training meetings would also make this sound in response to incorrect, misspelt or poorly expressed responses. Few laughed out loud during the standardising meetings, but some did; the Principal Examiner on English 1 was notable for her tendency to laughter, for example, which formed part of her ongoing commentary on the essays which she was reading aloud for her examiners.

The most usual affective reaction is a sympathetic one. In the 'think aloud' protocols, Lupin expressed repeatedly 'it's a shame' in response to the marks she was giving candidates, while being absolutely certain they were the right ones. However, it was not only sympathy which was seen in the examiners. They were more likely to indulge in empathetic reactions towards candidates, referencing their ability to conceive the external circumstances in which candidates sit exams, and suggesting a sense of fellow feeling for them. The constraints of time, or lack of constraints, are frequently referenced, as are the

demands of the papers. But it is not only this context that is specified. An extended discussion of one script by the Team Leader on History 2 to her team exemplifies this.

> She is really not off the ground. We could not squeeze her out. We gave her a 6. Scraped her into a 7 on the other one. She did refer to the sources. It is confused… but she does attempt to use the sources. Again if you gave her a 6 or a 5 the world is not going to end but be aware… you're in the area where she may fail. A couple of marks can make the difference between a fail**.** Twenty overall is unlikely to get her a pass – but in the AS overall it may make the difference. Could be the bit that doesn't get her into university… almost more difficult to mark. That's softie me.

Not only does this show the attempts to achieve the best possible outcome for the candidate, to 'squeeze' extra marks out of the script, but it also shows what seems to be one of the key aspects of the empathetic reactions to mental projections of candidates: the presentation of self. Here the team leader shows herself to be a 'softie' who finds this script more difficult to mark because it lies on the cusp of failing the candidate, a student who she will never know, but for whom she can feel sympathy. This tendency is reflected in the normal terms of reference of 'generous' and 'tight' or 'mean' marking. The discourse around marking is intrinsically emotive, and it is linked specifically to character: one Team Leader tells a member of her team that 'you are more generous. You're a nicer person' (English 1). Examiners are constantly told that they should err on the side of generosity; it *is* 'erring' yet in doing so examiners can present themselves as better human beings. It may not be coincidental that far more affect reactions were found in the live meetings than in the Voice Over Internet Protocol ones; face-to-face engagement with fellow markers may well prompt a greater awareness in examiners of what kind of person they appear to be.

Elsewhere participants showed empathetic reactions but also felt the need to demonstrate that they were not going to be biased by the emotion provoked by the script. One Senior Examiner remarked: 'I feel very sympathetic towards the candidate. That's just a comment, not anything to do with how we're going to mark it' (English 2). This is reflected elsewhere in the data when examiners are careful to show that they have identified potential sources of bias but will not be affected by them. They were also ready to demonstrate that inappropriately negative affective reactions were not directed at candidates: 'I don't like this man. That's the author, not the candidate' (Senior Examiner, English2). There was an awareness on the part of examiners that they might have emotional

reactions which might prejudice their judgement inappropriately. Lupin, for example, during her 'think aloud' recordings, while debating which mark to give an essay, commented:

> I think maybe I'm being unfair though, I have to make sure it's fair because sometimes you just take against an essay for no reason that you can really fathom, and that's not...

At this point she tails off, and notes another feature which leads her to think that she is 'being fair', in the terms of the strong concern expressed by many examiners at all stages of the process.

The examiners demonstrated a tendency to make deductions about candidates, typically making a distinction between the script and the candidate who had produced it: 'it's a more sophisticated candidate but not necessarily a more sophisticated answer' (Examiner, English 2). There is a sense in places that the script can in fact be detracting from a clear view of the 'true' ability of the candidate – so that one senior examiner commented that the 'florid' communication 'masks the fact that the candidate is not fully in control of the material' (Principal Examiner 1, History 1); this comment also reflects further the ability to distinguish between Quality of Written Communication and other skills where necessary, as discussed above.

However, the projections tend to go well beyond the simple acknowledgement of the existence of a candidate behind the script, or even simple speculation about what a candidate was thinking or doing in producing it. In particular, examiners show evidence of elaborate conceptions of how candidates produced their responses, constructing the instructions of teachers and the way in which students have learned the topic. Sometimes this is based on firm knowledge, as when some history examiners remarked on a number of candidates basing their answers around the debate in a major textbook instead of the one presented in the question. In other cases, they deduce teaching from the content of the essay, in a similar way to their deduction of other characteristics, so that we get: 'it's someone who's been told you have to move from one play to another all the way through the essay' (Examiner, English 2). In context this explained a slightly clumsy essay style, but there was not an explicit decision to make allowances for it because of its source. In contrast the same senior examiners criticised the naive assumptions expressed by one candidate about readers of a certain newspaper, but agreed that they shouldn't penalise them for that because 'it's what they're told' (Principal Examiner, English 1). In general the projections of teachers are in the candidate's favour, as they tend to construct poor performance as the

result of poor teaching, but, interestingly, good performance is seen as the student's achievement.

Can these affect reactions be seen in terms of the affect heuristic, then? Although there is plenty of evidence of affect responses on the part of the examiners, there is no evidence of these responses changing their decisions. On the contrary, affect reactions tend to be accompanied by hedging comments, which position the affect response as irrelevant to the judgement process. It seems likely that examiners' awareness of their affective responses to scripts mediates their potential as a source of bias. Zajonc (1980), however, would argue that it is an instinctive affective reaction which makes a decision, before the rational mind finds the reasons to justify it. It might be possible to argue that this is so in some of the decisions seen in the data, particularly those in Parrot's 'think aloud' data, where she tends to go straight to a mark, and then explain her decision. For the most part, though, examining judgements are mostly made in such a deliberated fashion that it is hard to see that they might be made instinctively, according to 'like', before a lengthy process of considering a wide variety of factors. It is possible that an affect reaction is pushing them in a certain direction, if not towards a specific mark. However, as with the care shown by principal examiners to defuse potential bias from cognitive heuristics, the hedging statements made by examiners suggests that by being aware of their reactions, they are able to counter them. This is not to say all affect reactions would lead to bias. It is clear that the very top level scripts stimulate an affective reaction of admiration, liking and general optimism: 'it's extraordinary, you just can't fault it, just certainly an incredible essay', as Parrot responds to one. Such use of superlatives in relation to the very top scripts is not unusual. Whether the affect reaction is driving the judgement, or the other way around, is impossible to tell.

Conclusions

It is clear that there is a role for cognitive heuristics in examination judgement; they are as applicable here as in other areas of human judgement, and can be seen in the real world as well as in the laboratory situations where they were theorised. Their impact is most obvious in the elements of the training meetings which relate to them, namely in terms of the selection of sample scripts. The concern of the principal examiners and the Awarding Bodies in selecting sample scripts is to mitigate potential bias which could come from the cognitive heuristics, even if it is not with conscious reference to *representativeness*, *availability* and *anchoring and adjustment*.

Of the three, availability is most clearly seen in terms of the sample scripts and potential bias. It is perhaps the most difficult heuristic to see in operation in the context of examiners' actual decision-making. However, there is extensive use of representative characteristics to help make judgements of scripts. They do not lead to complete decisions, however, rather being used as indicative measures of the appropriate range of bands in which a script should be located, so in the top two of five, or the middle three, or the bottom two. Characteristics which are representative are also used in the adjustment period of the decision-making. Anchoring and adjustment is seen as a conscious process in the examination of some modules, but it is also present in an automatic way when decisions are made, and is intrinsically linked with the comparison principle described in the previous chapter. The three cognitive heuristics posited by Kahneman *et al*. (1982) are clearly strongly inter-related in their application, and overlap to some extent. Most of the examples cited in this chapter can be interpreted as an application of more than one heuristic; it seems likely that all information which becomes a focus of examiners' attention can be used in whatever way is the most cognitively economical at the time. Some characteristics of essays, such as their Quality of Written Communication, or their level of analysis, may therefore be used twice in any given decision: both as a representative characteristic to locate a general area, and then more specifically to adjust a mark which has been suggested.

The unofficial heuristics which are provided at various points in the training process are designed to lead to accurate judgement, but with hedging to ensure they are not used too lightly. They can mostly be categorised as identification of further representative characteristics which examiners are aware that they are using. Awareness is clearly a key to ensuring that none of these potentially useful processes lead to bias, and much of the standardisation process seems to be designed to making explicit the judgement processes and their drivers both by and for examiners (see Chapter 4 above). Examiners are particularly aware of their emotive reactions to scripts, and know that affect reactions are an inappropriate influence on decision-making; they take steps to both make their emotions clear to themselves and to separate them from the judging process.

It cannot be certain that the examiners' awareness of any of these aspects prevents them from also operating unconsciously to drive decision-making. However, it does seem that by making them to some extent conscious thoughts, their most inappropriate application, which would lead to bias, might be avoided.

Chapter 9: Questionnaire results

"Not a merry thought comes to my mind without my being vexed at having produced it alone with no-one to offer it to." Michel Eyquem de Montaigne

This chapter reports the results of a web-based anonymous questionnaire carried out with a larger sample of examiners a year after the initial data collection period. Forty-five participants responded to the invitation to participate, but not all completed every item; one additional examiner completed section one only (and therefore omitted the questions on which subject and for how long they have examined[1]). The first section of questions contained statements taken from the 'think aloud' data, or paraphrases of them, and the third presented some of the initial, general conclusions which can be drawn from the first phase of data collection, and which were suggested in the preceding chapters. These two sections are divided into similar thematic sections, and will be considered under those thematic headings below. The intervening section of the questionnaire consisted of mainly open questions, about the training and marking process. The responses to these questions will be considered in a separate section at the end of the chapter.

The majority of the items in section one, which asked for simple yes/no answers, are reported for the whole sample. Of these thirty four items only seven had significantly differing proportions of yes or no by the examiners of different subjects on an independent samples t-test, and for only two of those was the majority answer different if the two groups were considered separately. These incidences are noted where they occur and discussed where necessary. Similarly, only one of the statements in the third section of the questionnaire had a significantly different response from history examiners as opposed to English, and this will be noted where it is relevant.

The intention in gathering data at the end of the January session was again to ensure that participants were 'reliable' examiners, who had been invited back to continue marking in the more select session. The number of years of A level marking most participants were able to claim was extraordinary: almost half, 20 out of 45, had more than twenty years experience, and just five had fewer than five years experience. Analysis of variance was calculated with years of experience as the group, and on only one statement

[1] This individual has been included in the counts which relate to the first section only; where cross-analyses or t-tests based on the demographic data were carried out, this participant's response is omitted.

was there found to be a statistically significant difference between groups. Indeed it was the only one that even approached significance. This will also be noted where it is relevant below. Otherwise, the level of examiner experience is not used to consider the responses.

It will be noted throughout that there are very few instances in which all respondents, or even the overwhelming majority of them, agreed. This is consistent with the impression that builds through all the data, that examining is a very individual process, which different examiners carry out in different ways. It is also notable that in many instances, particularly in the general items, statements which apparently contradict approved marking behaviours, often stimulate the greatest level of disagreement from participants, despite the fact that they are drawn from the data. It is noticeable in some of the responses to open questions that examiners demonstrate a level of defensiveness of the system, which is so often openly questioned by the media, teachers and politicians; it is possible that some level of defensive bias was operating in these examiners' answers. This is not in order to protect themselves, but rather the system.

Closed Questions

Heuristics

The formal heuristics covered in the questionnaire were *anchoring and adjustment*, *representativeness* and *affect*; the possibility of informal but more specific heuristics was also considered. It has been commented that identifying the unconscious cognitive heuristics by asking about them is problematic, both above and in Crisp (2008a). Examiners are used to reflexive thinking in adjusting their own marking, however, and the answers below suggest that where heuristics may lead to bias, they are aware of them.

Anchoring and Adjustment

The *anchoring and adjustment* heuristic is formally represented in at least the history examinations by the explicit instructions for examiners to find a band in which the script belonged, and then move up or down from the middle of the band. Although this was to some extent reflected in the responses to the generalised statements (Table 9.2) the more specific statements were more equivocal (Table 9.1). This may reflect the fact that the specific statements appeared to represent situations in which the heuristic might lead to bias, in their emphasis on prejudgment.

Table 9.1

(n.46)	Yes	No
"I can see from the beginning this is going to be a top band answer. Where does it belong in the band though?"	25	21*
"It looked as if it was better than this to start, but it's basically one long point so I'm going to put it down a bit."	23	23

Asterisked items (*) are those for which the subject of examiners is significant.

The first statement received almost opposite answers from the history and English examiners (English: Y=17, N=5; history: Y=8, N=15. *p*.003), although both had several on the minority side. There is a possibility that English examiners give greater weight to the opening of an essay, or that English essays are more likely to retain the same level of quality throughout than history answers. This split response was mirrored by the responses to the first of the generalised statements, which suggested examiners pick a specific mark early in the marking process and then adjust from that point. On that statement, history examiners were more likely than English examiners to pick an answer on the disagree side of the spectrum (*p*.038).

Table 9.2

(n.45)	Strongly Agree	Agree	Disagree	Strongly Disagree
"Examiners think of a mark early on and then adjust if up or down."	0	11	**29**	5*
"Examiners decide on a band and then decide on a mark within it."	20	**22**	2	1
"Examiners decide on a mark and then look at the mark either side to confirm."	3	**26**	15	1

Asterisked items (*) are those for which the subject of examiners is significant.

The remaining general statements were accepted by the majority of examiners, and the mandated procedure from History 1 which is reflected in the middle of the three statements received very wide support, despite some evidence from the 'think aloud' data that some examiners are more likely to go straight to a mark without choosing a band first.

Representativeness

Although examiners may disagree about the application of some representative characteristics, it is clear that some form of *representativeness* heuristic is acknowledged by

the majority of respondents. A general statement referring to a characteristic which is clearly unrelated to the quality of the work produced, good handwriting, is seen as representative by a very small number, but on the whole is rejected (Table 9.4). There is a sense of the 'construct' of different bands being recognisable to some examiners, which must draw to some extent on the representative characteristics. The length of essays is seen as being some kind of indicator, and shortness is not a representative characteristic for good scripts, either in the general or the specific statements.

Table 9.3

(n.46)	Yes	No
"Hmm this is very short. I don't hold out much hope for this."	18	**28***
"I have a kind of idea of what a band 3 looks like.	**25**	21
"It looks very short but the handwriting is very small. Very deceptive."	**32**	14
"This just feels like a top level script."	**27**	19
"I don't need to read the rest of this. It's clear from the first page it isn't going to get any better."	0	**46**[2]

Asterisked items (*) are those for which the experience of examiners (20+years) is significant.

However, there was a significantly different response from examiners with more than 20 years of experience in relation to the first specific statement given in Table 9.3 above. Experienced examiners were completely opposed to the statement, rejecting it by 17 to 3, in contrast to the majority 'yes' responses from the rest of the sample (Y=15, N=10), a significantly different response (*p*.002). (The different total sample numbers is due to the participant who completed the first section of statements but not the second section on their own background.) This suggests that extensive examining experience teaches markers that short answers can be very good; this speaks to the *availability* heuristic, as greater numbers of examples over time are presumably responsible for eroding this perceived representative characteristic.

[2] This was the item expected to prompt a 'no' from examiners.

Table 9.4

(n.45)	Strongly Agree	Agree	Disagree	Strongly Disagree
"Examiners have an idea in their head of what each level looks like."	6	**33**	5	0
"Short scripts do not tend to be top band."	1	**25**	17	2
"Good scripts usually have good handwriting."	0	3	**31**	11

Interestingly when asked whether they agreed that short scripts do not tend to be top band, it was precisely the more experienced examiners who agreed, although not in a statistically significant way. (It does become significant if examiners are categorised as 'experienced' after 10 years (*p.*024) but this puts 80% of the sample into that category.) This suggests perhaps that length is truly a representative characteristic, but only in certain ways for certain bands, and for other bands, it is an *availability* issue, as posited above.

Affect

Affect has been a minor concern of this thesis, as examiners appear to take steps to mediate its effect on decision-making (as shown in the second statement given in the table below). The phenomenon of the 'laughing rater' (Vaughan, 1992), however, and its correspondence in the 'tutting examiner' in this study, prompted the inclusion of these items in the questionnaire.

Table 9.5

(n.46)	Yes	No
"At last! Someone who's answered the question!"	**30**	16
"I feel really sorry for this kid. If I give this a five it's going to fail. But it's only worth five."	**29**	17†
"I think I can scrape this into the bottom of the next band up."	**29**	17†
"Oh why did you write that? You silly silly child.	10	36*

Asterisked items (*) are those for which the subject of examiners is significant.

† The resemblance between these two results is coincidental. Only 20 examiners said yes to both.

The responses demonstrated that a range of affect reactions are experienced by examiners, although the impatience of the final statement is not reflected by a majority; perhaps

sympathetic affect is more compatible with examiners' desire to present themselves more positively, discussed in the previous chapter. (The difference in response between subjects was only in the degree of rejection; English examiners almost unanimously rejected the statement.) The substantial agreement with the first three statements in Table 9.5 above suggest that examiners engage affectively both with the process, and with representations of the students involved, a theme which is explored under 'deduction' below.

Informal Heuristics

The statements, both general and specific, which spoke to the theme of informal heuristics were among the most widely rejected of the questionnaire data. All had at least some agreement but for some it was extremely weak. The high negative response to the first statement in Table 9.6 is a positive result, in terms of the marking process, as grade boundaries change frequently and a numerical marker is not a safe reference point. The second specific statement has only marginally more respondents in the 'yes' category than the 'no', but demonstrates that 'editing' of the mark scheme criteria occurs in some instances, with certain ones being seen as more predictive than others, so more worthy of a focus of attention.

Table 9.6

(n.46)	Yes	No
"I know we use bands, but I know where the grade boundary is and I use that as a reference point."	7	**39**
"My team leader told me that this factor was the most important in choosing a band."	**24**	22

The survey respondents did not perceive examiners as likely to use informal heuristics, however, which suggests that the results shown in Table 9.6 are not due to the selection of statements. The first of the general statements in Table 9.7 could also be said to refer to the cognitive heuristics which aid decision-making. The term 'rules of thumb' is potentially distracting, in that it might be read as being pejorative by the participants, who, as noted before, were defensive of the examination marking system. There is no good alternative, however, since 'heuristic' is not a term in common use. It is also in conflict with the very positive agreement with the statement 'Quality of communication is a good indicator of how good the student is' (see Table 9.16 below), which in itself is a rule of thumb.

Table 9.7

(n.45)	Strongly Agree	Agree	Disagree	Strongly Disagree
"Examiners use rules of thumb to decide on how good scripts are."	1	16	**21**	7
"Examiners ask their team leaders for a rule about, for example, how many quotations are needed to reach a certain level."	0	14	**17**	13
"Examiners decide for themselves which factors in the mark scheme are most important."	0	3	**28**	14

More interesting is the almost complete rejection of the final statement regarding examiners' own filtering of the criteria given in the mark scheme. This suggests that very few examiners consciously make judgements about which criteria seem to be dominating the judgements of senior examiners, or that a hierarchy of importance, with some criteria emphasised, is established for them during the training, so that they do not need to make that discrimination for themselves. The latter is partially supported by the fact that half of participants recognised the statement concerning a team leader highlighting the importance of any given factor.

Comparison

The use of comparison in making judgements of papers was one of the most strongly supported principles in the questionnaire. Seven statements which drew on the principle of comparison were given in the first section, four of which referred to the sample scripts, which represent the sanctioned touchstones for comparison, although their systematic use as such is not suggested by the rubric. The three statements which suggested that comparison was used in the case of difficulty, or in order to check a decision which they had tentatively made, gained overwhelming support, in the order of 3 to 1. (Marked with + in Table 9.8 below.)

In contrast, the fourth statement relating to the sample scripts, that they were used 'all the way through the marking session' was exactly divided between positive and negative responses. The use of an 'imagined touchstone', in the terminology suggested in Chapter 7 above, also received a positive response. Perhaps the fact that the sample scripts are not seen to be used all the way through the marking period reflects the fact that there is other potential comparative material, or perhaps comparison becomes less used as examiners

become more practised (as was suggested by the 'think aloud' data in the earlier part of the study).

Table 9.8

(n.46)	Yes	No
"So I think it's 25. Let's check it against the sample scripts." +	**30**	16
"A better script would have thought about context too."	**35**	11[3]
"I use the sample scripts all the way through the marking session."	23	23
"Not sure about this one. How does it compare to the sample scripts?" +	**37**	9
"Right, okay. So how does it compare to the sample scripts? It's better than M. Not as good as T. That would make it Band 4. So.... 18 marks." +	**32**	14
"That makes 26. But it wasn't as good as the last one. Hmm. Make it 24."	11	**35**
"This second essay is stronger than the first. Did I miss something?"	**28**	18

The final two statements suggest the use of comparison between actual live scripts, one which suggests changing in favour of the candidate, which was accepted, though not as strongly as some of the other statements, and one which suggests adjusting down by comparison with the previous essay, which was rejected, again in the ratio of 3:1, by respondents. The one which was rejected also represented automatic adjustment without any thought, rather than the careful reconsideration suggested by 'Did I miss something?', which might also have stimulated the rejection. It must also be considered that the 'Did I miss something?' statement is one of those which received significantly different results from history and English examiners: 18 of the English examiners accepted it with 4 opposed, but a small majority of the historians answered 'no' (13 in contrast to 10 acceptances) (*p*.007). It is possible that this is due to the representativeness heuristic, and that history examiners are more used to seeing two essays in the same script which are at different levels (although this directly contradicts an informal heuristic given by the Principal Examiner on History 1). Given the strength of the positive responses to the statements in the

[3] This item was inadvertently included twice in the questionnaire. The results given are the first set of responses. The second time the item appeared four participants chose 'no' instead of 'yes', giving a 2:1 ratio instead of 3:1, because of the relatively small sample. It is still clearly an overall agreement.

first section relating to comparison, the responses to the general statements which explicitly used the word 'comparison' were unsurprising.

Table 9.9

(n.45)	Strongly Agree	Agree	Disagree	Strongly Disagree
"Examiners use comparison with other scripts they've just marked to decide on a mark for a script."	3	**26**	14	2
"Examiners use comparison with sample scripts to decide on a mark for a script."	9	**32**	2	1
"Examiners use comparison with their idea of what a good essay should be to decide on a mark for a script."	1	8	**30**	6

It is clear from these results that comparison with tangible material is a well-acknowledged practice during marking (although the confident agreement with the use of other live scripts is worrying given the potential for the halo effect to cause bias). The imagined script, which received strong support when phrased as a thought which examiners might recognise, does not receive similar support as a general principle; rather the opposite. The principle does of course include the phrase 'to decide on a mark for a script'; there is plenty of evidence in the data to show that examiners make mental notes of characteristics which they claim do not affect their decisions. Nonetheless there is sufficient agreement with the statement to support the inclusion of the imagined script among the potential touchstones for examiners' use, and to some extent to support Pollitt's (2010) claim that where material is not explicitly provided for comparison, examiners will use mental images of them.

Self-adjustment

Self-adjustment, though dealt with separately here, is an informal heuristic in itself, in that examiners use the knowledge of their own standard as a rule of thumb to adjust their original decision to ensure it is aligned with that of the principal examiner. It is not a heuristic which is given to them, but nonetheless it is applied to alter their decision-making, if not their judgement of the script. The two statements given under this category were a contrast pair, to which it was expected that examiners would answer 'yes' to a maximum of one. Despite this two examiners did in fact answer yes to both, as can be seen in the cross-tabulation.

Table 9.10

		I was too generous on the training scripts so I need to choose the lower of these two marks.		
		Yes	No	Total
I know I tend to be harsh, so I need to give them the benefit of the doubt and bump it up a mark.	Yes	2	5	7
	No	12	25	37
Total		14	30	44

Twenty five examiners (more than half the sample) said no to both statements, but seventeen said yes to one of the two. Such a systematic adjustment suggests that for a good proportion of examiners, the process of bringing their standard closer to the approved one is a very conscious process, and one which is very regular. It supports the idea that examiners engage in meta-cognitive evaluation during their marking, and relates to the theme of agency found in the analysis of the open question responses below.

Deduction

One of the features of examiner thinking which emerged during the first phase of data collection was not strictly related to decision-making, but was related to the affect reactions which examiners had to scripts and to candidates, or rather to imagined representations of candidates (described in the previous chapter).

Table 9.11

(n.46)	Yes	No
"He's working through a checklist he's been given."	**36**	10
"Must be a girl with that handwriting."	19	**27**
"This candidate has been taught really badly."	**31**	15
"This candidate should have done better. There's potential but you're not marking that. You're marking what's there."	**37**	9

Examiners demonstrated deductive reasoning in constructing images of the students behind the answers. Three of the statements which were taken from the data relating to deductive reasoning received a large majority of positive responses (between 67 and 80%). The fourth, relating to the suggestion of gender by handwriting, was more frequently answered in the

negative, but still had 40% positive responses. This is a very high level of recognition by the participants, and among the most decisive results from the questionnaire.

The general statements also included a section on deduction:

Table 9.12

(n.45)	Strongly Agree	Agree	Disagree	Strongly Disagree
"Examiners can tell things about students' characteristics from what they write.	0	13	**17**	15
"Examiners deduce things about how a candidate has been taught from what they write."	6	**30**	8	1
"Examiners can see if a candidate has not done as well as they should from what they write."	0	**21**	**21**	3

Considering these general statements in the light of the yes/no items from section one suggests some intrinsic contradictions: in particular the emphatic disagreement with the statement that 'Examiners can tell things about students' characteristics from what they write', which contrasts with the strong agreement with the final yes/no item ('This candidate should have done better...'), which would also suggest a stronger response on the 'agree' side of the scale for 'Examiners can see if a candidate has not done as well as they should have from what they write.' A cross comparison of responses to these items showed that 11 of those who recognised the statement 'Must be a girl with that handwriting' answered either 'disagree' or 'strongly disagree' to 'Examiners can tell things about students' characteristics from what they write.' For the same general statement 24 of those who agree with the final yes/no item, 'This candidate should have done better' chose either 'disagree' or 'strongly disagree'. Similarly, four of those who recognised the statement 'he's working through a checklist he's been given', disagreed with the general statement that 'examiners deduce things about how a candidate has been taught from what they write.'

These discrepancies may suggest two things. Firstly, it may be that examiners think differently during marking than they believe they do, so that asking them about how they make decisions is an ineffective methodology (which is part of the reasoning behind constructing the 'think aloud' method, and analysing real data from standardising meetings). Secondly, it may be that their responses on the general statements are coloured by what they perceive the official system should be, as is consistent with the somewhat defensive tone of the comments discussed under 'Open Questions' below. The belief in being able to

identify gender from handwriting, for example, is impolitic in examining, where the anonymity of the candidate is intended to be integral to the fairness of the assessment; where examiners make assumptions about a student's gender in the phase one data, it is almost always accompanied by some form of apology or hedging statement, as they acknowledge it is inappropriate.

Quality of Written Communication

Nominally Quality of Written Communication (QWC) is a marginal but nebulous area of the mark scheme, usually remaining unspecified but with a limited influence over the judgement as a whole. The evidence from the 'think aloud' data and the standardising meetings suggested that QWC was much more influential in examiners' decisions, and was often regarded as both representative and also as a potentially deceptive indicator. The specific statements which were chosen reflected the second aspect, in that they described moments when the QWC was at odds with the contents of the essay. They all, however, implied a judgement was made of the writing quality, whether or not it was in contrast to the assessment of the other characteristics.

Table 9.13

(n.46)	Yes	No
"The content is okay, but the style's a bit weak."	**34**	12
"The style of writing makes it seem better than it is."	**32**	14*
"The quality of written communication is terrible – the content is much better than it appears."	**37**	9*

Asterisked items (*) are those for which the subject of examiners is significant.

The statements in Table 9.13 received strong support from the respondents, particularly the examiners of history (representing more than a quarter (two out of seven) of the instances in which examiner subject was statistically significant). The difference in response patterns for the second two statements is given below.

Table 9.14

	History		English		
	Yes	No	Yes	No	*p.*
"The style of writing makes it seem better than it is."	20	3	11	11	.007
"The quality of written communication is terrible – the content is much better than it appears."	15	8	21	1	.010

English teachers are more likely to see bad QWC as deceptive; good QWC may be deceptive for history examiners, but the response from English examiners on the first statement is more ambivalent. Potentially for English examinations the two different aspects are not as separate as they are for history, so that it does not makes sense for the writing to be better than the content, although the reverse is apparently possible. History examiners are quite clearly more likely to consider the QWC to be a potentially distracting factor in making an accurate judgement of the quality of an essay.

Table 9.15

	Strongly Agree	Agree	Disagree	Strongly Disagree
"Quality of written communication is a good indicator of how good the student is." (n.44)	2	**28**	13	1
"Examiners only notice writing style when it's very good or very bad." (n. 45)	2	15	**24**	4
"How students write affects the mark they receive even if it's not explicitly in the mark scheme." (n.43)	3	**23**	15	2

The disagreement to the second statement in Table 9.15 suggests that examiners are well used to making judgements of the quality of writing, no matter at what level that judgement falls, and that the QWC is a legitimate and frequent focus of attention. Its position as both a representative indicator and a characteristic which affects the decision which is made is quite clearly established by the majority of positive responses to the other two statements. Examiners also indicate their willingness to assess outside the mark scheme in their response to the final statement in Table 9.15. The evidence from the questionnaire

undoubtedly supports the suggestion that Quality of Written Communication plays a larger role in examiners' thinking than its official mandate.

Marking behaviour

These statements constitute a mixed group, which refer both to behaviours and to the way in which examiners relate to the mark scheme. The first three statements are somewhat interrelated. Both of the first two relate in some way to the idea that decisions become progressively more automated, and potentially to the theory of System 1 judgements migrating to System 2. Exactly half the examiners agreed that the decision-making process became easier, yet only three of those who found it becoming easier felt able to abandon the use of the mark scheme altogether, perhaps indicating the strength of the emphasis placed on it during training. One of these three, however, is a Team Leader (as shown by their answers to the open questions), and all three are very experienced examiners, with 15, 18 and 27 years of experience; clearly this trait has not had a noticeable impact on their decision-making. All three also answered 'no' to the statement about having the mark scheme in front of them; only one other examiner joined them. More than 90% of the respondents agreed with the statement that they had the mark scheme in front of them at all times. This might seem to contradict the basis of this study, but it is undeniable that the mark scheme is the nominal centre of all marking activity, despite the evidence that examiners use other mechanisms to assist them in making reliable decisions; it is prudent to keep the mark scheme to hand, and doing so is emphasised as good marking behaviour. Having it available does not mean it is being used systematically for every decision. A further statement, which was included in the 'Comparison' theme above, also speaks to marking behaviour: 'I use the sample scripts all the way through the marking session' received equal numbers of positive and negative responses, demonstrating that they are not considered to be as important as the mark scheme (or as desirable in the eyes of the system, perhaps), despite their apparent usefulness.

The first statement, that it becomes easier to make a decision, was one of only two statements in the whole section to stimulate different majority responses from the two subjects. Fifteen English examiners agreed, as opposed to only 8 historians, with 7 and 15 choosing 'no', respectively (*p*.025); this is interesting in the light of the statement's source, the Team Leader in History 2. This may suggest the migration of marking from conscious and conscientious decision-making into a more automated process is more likely for English examiners, although both subjects demonstrate it.

Table 9.16

	Yes	No
"It gets easier to make a decision. I mean, fifty scripts down the line, it's more intuitive." (n.46) *	23	23
"By the time I'm marking the text scripts I don't need to use the mark scheme at all." (n.46)	3	43
"I have the mark scheme in front of me at all times." (n.46)	42	4
"I've not usually got a firm figure in my mind the first time I read through.... then the second time I make a decision." (n.45)	17	28
"Is this some awareness? Or is it limited awareness?" (n.46) *	32	14

Asterisked items (*) are those for which the subject of examiners is significant.

It is also English examiners who are more likely to recognise the statement which queries what the concrete representation of an abstract criterion of the mark scheme looks like: 'Is this some awareness?' Although more participants answered yes than no to that final item in this table, it is marginal in history (12 versus 11), but overwhelming in English (20 versus 2). This is a statistically significant difference (*p*.003). Potentially it is the wording of the question which has caused this difference; 'awareness' is a term which is more likely to feature on an English mark scheme than one for a history module. It is also possible that English examiners are more nervous of the potential for misunderstandings and errors in examination marking, given the controversy around English and assessment which was discussed in Chapter 2, which has no equivalent for history teaching.

A substantial portion of participants, 37%, agreed with the statement that suggests they commonly read an essay more than once before reaching a mark; something which substantiates the idea that marking is a challenging cognitive activity. It is surprising, however, given the time constraints involved, that examiners have enough time to read even most essays more than once, unless it is that one is merely a skim reading, either to identify the main points of interest for greater detail, or to confirm the presence of features already established, but not fully remembered. The latter was more likely for the three examiners who provided 'think aloud' protocols for the main study, who often 'flicked' back through an essay to confirm their impressions and quantify a judgement.

Open Questions

The open questions spoke more broadly to the examining process and in particular to the training. Rather than report on the complete responses to each question, the discussion here will draw out elements of examiners' answers which relate directly to either the cognitive processes they engage in or the conclusions drawn from the data within this thesis. Some of the questions addressed the format of the training meeting; while only one participant worked on a module which was based on online training, some others had had experience of such standardising. All the participants preferred face-to-face marking, and their varying explanations refer to many of the elements which have been mentioned in previous chapters. There has been controversy among examiners about the gradual move from live to online training meetings: a survey carried out by one of the major exam boards on their examiners last year showed that 38% of their respondents were not happy with the online training system (Edexcel, 2010). This is a specific system, as opposed to the concept, but it is related to the question of online versus face-to-face training.

The idea that face-to-face training enabled the creation of a sense of team identity, or collegiality, was raised by five examiners; the complementary point that online training was 'depersonalised' was raised by several more. I suggested in Chapter 4 that the creation of a team identity was in some cases a deliberate training strategy, and several of the participants who identified themselves as team leaders or principal examiners in their answers to the open questions suggested that this creation of a team in a face-to-face situation made it much easier to identify examiners who were struggling or insecure, and to gain an accurate gauge of their accuracy in context. Perhaps team leaders use comparison to judge the effectiveness of examiners just as examiners use comparison to judge scripts. The dominant reason for preferring face-to-face meetings, mentioned by more than half the examiners, was the speed and ease of the discussion in a live context, both with supervisors and with other examiners. This was intimately connected with some of the reasons examiners gave for marking, in that professional development, intellectual curiosity and connection with colleagues are all felt to be stimulated by face-to-face conversations.

In this, and throughout the open questions, examiners demonstrated a sense of agency in familiarising themselves with the mark scheme and the standard which has not been seen to any great extent in the data in this study. One stated: 'I like to be able to query and argue until I am confident with the mark scheme', and this sense of interactivity was important to several participants. This is commensurate with the approach which is taken in standardising meetings: it is not a case of learning the mark scheme, but of learning to

'interpret' it as another survey respondent phrased it. The sense that marking is not a simple task was also conveyed by one comment that mentioned the 'many nuances' of the judgement process. Beyond that, some mentioned the 'stress' of the marking period. One examiner said, on learning the mark scheme, that it was 'always something of a struggle and always accompanied by the feeling that this time I'll be found out.' This sense of a combative aspect to the process, or a fear of the supervisory process is an interesting element; it suggests the hierarchy can even be threatening to the lower members of the group (who can be stopped from marking). It hints at another affective aspect to the process.

All but one of the examiners in response to the question 'how important is the discussion between members of your team?' included a variant on 'very', 'essential', 'crucial' or 'critical', with most simply choosing the first. Two longer comments (both by English examiners) began in the same way and then went on to make contrasting but illuminating comments:

> 'Very important: I try to make sure that everyone has their say. There is a "bottom line", though: our task is to mark to the standards defined by the Principal Examiner.'
>
> 'Very: in fact, it's vital in defining the standard: in fact, the discussion is the standard. It doesn't exist as an abstract phenomenon. We discuss to establish what we want, and that cannot be entirely defined by the mark scheme: the discussion is part of the standard.'

The former approach (summarised by a self-identified team leader) is clearly that favoured by Awarding Bodies and OfQual; both comments reflect the fact that the mark scheme itself is not enough. For one examiner, it is the Principal Examiner who embodies the standard (presumably conveyed jointly through the training and the sample scripts), and for the other the discussion. A further comment from a history team leader suggested that discussion for her leads towards the 'agreed standard' as she can 'talk them through why the mark was awarded and they can explain why they gave the mark they did.'

Additionally, there was a sense of pride in the system, and a desire to defend it from criticism, evident in the answers to the open questions. One question asked if examiners thought their students got the right grade in the examination, and for them to expand on that. No formal analysis of the answer to the yes/no question was done since it was clear that many examiners had misinterpreted the question as a challenge to their own marking, by considering 'your students' to refer to the candidates whose work they were marking.

The answers to the open question, however, produced both by those who understood and those who misunderstood the question, were often quite sharp or emotional. One examiner stated 'I have real faith in the way in which we – on behalf of the Board – carry out the examining process.' Others used the words 'rigorous' or 'great care' to describe the approach of the system. One cited the system of 'checks and balances in History' which 'generally ensure that errors are identified and indifferent markers weeded out.' Questions about the mark scheme, or the training method also occasionally stimulated statements of a similar quality, and the final question, which asked if they liked examining, often conveyed a sense of job satisfaction and pride in a 'worthwhile' process, as well as the 'enormous respect' for their colleagues.

Examiners were not unanimous in their praise for the system, however. One stated that she thought 'the general quality stays high – despite the general assault from the press/ politicians and, unfortunately such exams as the Pre-U', but nonetheless warned that marking 'is not, however, a science, nor can it be in the humanities.' She thus identified one potential cause for the defensive attitude in some comments, but acknowledged that the precision which some examiners wished to claim for the system was a false one. It is also a system made up of subjective judgements, as one examiner commented, in answer to the question on whether her students got the right grade:

> Sometimes. English Literature is an interpretative subject. I know that everything is done to ensure standardisation of marking through an omnicompetent Principal Examiner managing a rational process effectively, but you still, occasionally, get unlucky with an examiner.

While I might take issue with the full extent of the rationality of the process, it is clear that some examiners understand the limitations of the system, while continuing to do their best within them. The word 'professional' was mentioned by eight of the respondents to the open questions; it seems that for them examining is an important, respected job which is deserving of their full care.

Conclusions

The response to the questionnaire confirms the impression created by the first stage of the study, which is that there is no single way in which all examiners make decisions, even the mandated way. A wide variety of factors influence judgements, which are made in a number

of ways. Among the most widely used and recognised cognitive strategies are *anchoring and adjustment* and comparison.

A table (9.16) was constructed containing all the general items from the questionnaire, with rankings calculated by awarding 2 points for each 'strongly agree' through to -2 points for 'strongly disagree', weighted around zero. No consecutively ranked statements came from the same group. Four themes had two statements each with positive residuals: anchoring and adjustment; Quality of Written Communication; representativeness and comparison. Neither the section on informal heuristics nor on deduction had any statements which had an overall positive response, despite the fact that the specific statements relating to those themes had at least some positive responses for each.

Table 9.17

Examiners decide on a band and then decide on a mark within it.	58
Examiners use comparison with sample scripts to decide on a mark for a script.	46
Examiners have an idea in their heads of what each level looks like.	40
Examiners deduce things about how a student has been taught from what they write.	32
Quality of written communication is a good indicator of how good the student is.	17
Examiners decide on a mark and then look at the mark either side to confirm.	15
Examiners use comparison with other scripts they've just marked to decide on a mark for a script.	14
How students write affects the mark they receive even if it's not explicitly in the mark scheme.	10
Short scripts do not tend to be top band.	6
Examiners can see if a candidate has not done as well as they should from what they write.	-6
Examiners only notice writing style when it is very good or very bad.	-13
Examiners use rules of thumb to decide on how good scripts are.	-17
Examiners think of a mark early on and then adjust it up or down.	-28
Examiners ask their team leaders for a rule about, for example, how many quotations are needed to reach a certain level.	-29
Examiners use comparison with their idea of what a good essay should be to decide on a mark for a script.	-32
Examiners can tell things about students' characteristics from what they write.	-34
Good scripts usually have good handwriting.	-50
Examiners decide for themselves which factors in the mark scheme are most important.	-56

Some patterns in answers can be seen. The items which look like they might be a cause of bias tend to get negative answers, possibly because the sample is composed of experienced examiners, or potentially responses were influenced by an awareness of what the 'correct' answer should be causing editing of personal reactions. However, certain items, particularly relating to representativeness, suggest that examiners are in fact aware of

instinctive reactions which can potentially skew their judgement, and have made a note that certain characteristics are not representative; this would be a part of the learning process, like the fact that more experienced examiners do not necessarily think that short scripts are bad scripts.

This relates to the impression of agency which is given by the responses to the open questions. Examiners are not unthinking automatons who are trained by rote: it is an interactive process in which they are engaged. The examiners are also invested in the system: there is a defensiveness of the results produced by it, and an intention to produce the fairest outcome for students.

Importantly, all propositions presented in the questionnaire received at least some measure of support from the examiners who responded. There were relatively few differences related to subject or to length of experience. Examiners simply have different ideas about how examining works, and work in different ways. To judge from contradictions between responses to specific and general statements, examiners may also not have a clear idea of their own working.

Chapter 10: Conclusions

> *"I take it we are all in complete agreement on the decision here? Then I propose we postpone further discussion of this matter until our next meeting to give ourselves time to develop disagreement and perhaps gain some understanding of what the decision is all about." Alfred P. Sloan, Chair of General Motors*

The previous six chapters have explored a number of aspects of the process of decision-making in examining A level essays in the subjects of English and history. If any over-riding theme has emerged, it is that there is a great variety in the approaches, mechanisms, and foci of attention which can be demonstrated by examiners' decision-making. This chapter will explore some of the conclusions which can be drawn, and the aspects of the marking process and training which have become apparent, linking them to some of the literature explored in Chapters 1 and 2. Initially the research questions will be considered, outlining the findings that relate to each in a brief summary. The model of the process proposed by Crisp (2010a) will be considered in the light of the data. Some more general conclusions arising from the data, and their relationship to the theoretical framework of Chapter 1 will then be discussed, aiming to characterise the nature of the decision-making processes involved in the examination of A level History and English.

How do examiners make their decisions when assigning marks to essay scripts?

What decision-making behaviours do examiners exhibit?

Examiners exhibited a range of decision-making behaviours. The most consistently demonstrated behaviour was the use of comparison, which was carried out with any material which was available, and sometimes with mental depictions if suitable comparative material was not available. Examiners demonstrated the ability to remember what would be appropriate touchstones from the scripts which they had seen, even into the main marking period. The use of *representative* characteristics and the process of *anchoring and adjusting* spoke to the presence of the cognitive heuristics, but the presence of *availability* was most suggested by the selection of training scripts by the principal examiners, rather than in the behaviours of examiners. One examiner suggested that she worked down the mark scheme from the top to locate a script at the appropriate level: this is a behaviour which is entirely

rule-based and does not conform to the anchoring and adjustment procedure suggested by senior examiners.

The use of balance to adjudicate the level at which a script should be placed was notable in the discourse of examiners, but not in that of senior examiners. Comments weighted good features against poor ones. Balance was a particular feature of the summative comments of examiners, both during training meetings and during the 'think aloud' recordings. Additionally examiners demonstrated checking behaviour after they had made a decision, either with the use of comparison, or consideration of the numeric total against the holistic impression, or with the use of the mark scheme. Checking occasionally took place before decision-making also, if a conflict had arisen between two aspects of an essay: the quality of the conclusion in comparison to the impression created by the introduction, for example. Difficult decisions were associated with a greater period of checking, the use of a greater range of tools, particularly in that the mark scheme was more likely to be utilised, and also with a longer period of time to make the decision. They were also more likely to be associated with examiners going back to look at the essay a second time. In addition difficult decisions, or post-decision checking, might require revisiting a previous essay and altering the mark of either in comparison.

Examiners also engaged in hedging behaviours around topics or foci of attention which they did not consider to be appropriate drivers for decision-making. These included affective responses. This is not strictly speaking a decision-making behaviour, but it is one which affects decision-making; it has been suggested in the course of the thesis that their self-awareness and hedging was a mechanism which mitigated the potential effects of bias.

What training strategies are demonstrated by senior examiners and how do they relate to the decision-making behaviours exhibited?

A range of training strategies were outlined in Chapters 4 and 5. These included the use of model analysis of scripts by principal examiners and team leaders, which was partly demonstrated by junior examiners. The noting of specific details to create an 'outline' of the essay was seen both in the model analysis and in the practice of individual examiners.

In some meetings the sample and standardising scripts were used to create a framework for the range of potential responses and awards; some principal examiners deliberately ordered their scripts, although others did not. It was common for the anchor scripts to give examples of both the top and bottom ends of the achievement range, to set the 'parameters' (TL, H2), and to give multiple examples of the top level. The selection of

scripts demonstrated a concern to mitigate potential bias due to the *availability* and the *representativeness* heuristics, although this was not explicitly stated; it is more likely to be a response to actual past instances of bias than to the cognitive theory. This strategy provided examiners with a framework into which other scripts could be fitted; this relates both to the idea of a mental framework which is discussed later in this chapter, and to the extensive use of comparison by examiners.

In addition senior examiners modelled and demanded the use of anchoring and adjustment as a technique for reaching a mark: scripts were located within a band before various characteristics were used to decide on a final mark. To some extent this was also seen in the decisions which the three examiners made in their 'think aloud' recordings. As with the use of the mark scheme, as training progressed, and as the marking period progressed, the frequency of conscious use of this technique became substantially lower. The senior examiners also placed less emphasis on the detail of the mark scheme as the standardisation meeting progressed.

Other training strategies related to the interpersonal aspects of the examining process rather than directly to decision-making behaviours. These were seen to contribute to the creation of a productive working environment and a good working relationship. They were also related to the emphasis on professionalism, which was a recurring theme of the training. This may not relate directly to decision-making but it promotes continued care and attention on the part of the examiner, and encourages the most effective method of monitoring behaviours, which is by the examiners themselves.

What foci of examiners' attention during reading can be identified?

Examiners' attention can focus on a very wide range of factors during marking; the most frequently recurring foci included the specific focus of the module (the 'stated factor', the 'sources' or the 'context'), the discourse markers which candidates used to signpost their answers, the structure of the essay, the Quality of Written Communication, the 'approach' of the candidate, the analytical/ descriptive qualities and the obvious errors. Attention could also be drawn to the length, to the handwriting, to the perceptions of the candidate and the teaching which they had experienced. In addition examiners were aware of themselves during the reading process, and demonstrated on occasion that they were monitoring their own responses concurrently with their reading of the essay, so that their own reactions became a focus of attention.

The foci did not usually remain the same from essay to essay, although some examiners were associated with specific concerns. Occasionally examiners, particularly senior examiners, would deliberately state that a factor which had been the focus of attention was not a factor in their decision-making.

What tools do they use to make their decision and what is the role of the mark scheme?

Examiners will use any tool which they have at their disposal when they need to: the mark scheme, the standardisation scripts and other, live, scripts which they have marked, as well as their 'notes' from training, their personal experience of teaching the module, or knowledge of the relevant study texts, as well as their supervisors and their fellow markers during the standardisation meeting. The only tool which was provided but left for the most part unused was the indicative content.

The mark scheme, though nominally the centre of the training and the marking process, was not as dominant in either the meetings or the 'think aloud' recordings as might have been expected. It was particularly used at the beginning of the training, and was associated with resolving difficult decisions at some points, or in post hoc confirmatory checks. It might be used to 'fine-tune' a decision after a heuristic had guided the examiner to an area.

How do examiners make their decisions when assigning marks to essay scripts?

The conclusions which can be drawn from the findings of the sub-questions is that examiners make their decisions in a number of ways, drawing on a range of tools, and paying attention to different things. They will select the most relevant or the most easily available factors for a given script and the most relevant comparative material; the ease or difficulty of the decision will dictate which, or how many, of the decision-making behaviours they use, and which or how many tools are applied. They make their decisions with a degree of self-awareness as to their instinctive responses, and claim to adjust accordingly.

A potential model

Crisp's framework (2010a) suggested a model for the judgement processes involved in examining which is reproduced on page 41 of this thesis; the model divided the process into three phases, with an additional prologue and epilogue. The decisions demonstrated in the data from the 'think aloud' recordings from this study support this division of the process, and the stages which she has elaborated are consistent with the procedural elements of the

decision-making which Lupin, Caspar and Parrot demonstrated. It is a sufficiently general model to be applicable in a wide range of subjects and examination types. It does not, however, reach the full range of the cognitive processes which were identifiable in the data from this study. In particular, there is a concern with checking decisions demonstrated, both in training meetings and by individual examiners; this would be namelessly subsumed into the 'epilogue' of Crisp's model. Although it was not present for all scripts (as she herself admits for the stages in her model), it is certainly common enough, and substantial enough to require inclusion as a separate stage in the process.

Crisp's model is, it must be said, a model of the processes, rather than the cognition involved. Although some exemplar cognitive behaviours are included for some levels, and some psychological theory was integrated into the discussion of the model, it does not cover the full range of potential contributory factors. In particular, the evidence of this study would suggest that comparison must also be a specifically acknowledged decision-making behaviour, which might occur at any of the stages, but which is most likely to occur in the 'Mark Decision' of Phase 3.

As Crisp did, I have found evidence for the presence and use of all three of the cognitive heuristics posited by Kahneman *et al*. (1982, revisited in Gilovich *et al*., 2002) and it seems clear that models of the decision-making processes of examiners must include these. While the pragmatic framework of Graesser *et al*. (1997) has been of use in understanding the responses of examiners, it is too complex to be of direct use in modelling their decision-making processes.

The focus on judgement

The emphasis in the system of national examinations in the UK, from OfQual down, in the procedures of the Awarding Bodies, in the training and monitoring of examiners, is to control the process, in order to remove as much variability as possible. But this emphasis does not attempt to reduce markers to automatons; the focus is on judgement, rather than pattern matching or other more simple cognitive processes. The training emphasises the need for examiners to be alert to their own thoughts, and to the process; the motivational and interpersonal aspect of the training aims to promote their confidence, without leading to over-confidence. Examiners are asked throughout the process to explain the thinking which has led them to a particular mark; they are encouraged to consider each essay within the context of its production, and to explore their reactions to it.

Similarly, the analysis and commentary on scripts seeks to provoke thought among examiners rather than simple recognition. Once a feature has been identified, it is explored and discussed: is it an indicator of the presence of something else; how well has it been done; is it part of the requirements of the assessment objectives? The need to weigh up and consider the different aspects of a script rather than merely search for and identify features is an acknowledgement of the subjective nature of the process. Despite the desire for, or perception of, examination as being a matter of measurement, within the system it is clear that there is an understanding that it cannot be a simple matter, and that it requires professional judgement to be put into operation.

It was noted at the very beginning of this dissertation that the documentation suggests that OfQual prioritises 'consistency'. The question arises: is the process of examination consistent? The evidence presented in this thesis is that examiners' decision-making is anything but consistent in terms of the process; a wide range of tools and processes have been described here, and they are all demonstrated by the majority of examiners, at some time, but they are not all used all the time. The only consistency is in the outcome: to return to the nature of the sample, the examiners who took part in this study were all highly experienced, and had all been invited to mark in the January sessions, which gives some guarantee of their reliability and their ability to make correct decisions, however they make them. The process is not consistent, except where it matters.

However, when that is the case, the organisation relying on this professional judgement has a difficulty. Some examiners may make the correct decisions instinctively, and may do so for their entire examining careers. But if decision-making is left at the instinctive level, then those who do not make the correct decisions cannot be trained into better ones, and if a 'good' examiner suddenly becomes poor, unless there is some understanding, on her part and on the part of her supervisors, of the processes she uses to reach her decisions, then she cannot be helped to return to her previous reliability. The process must be conscious, but individuals do not apply mechanisms or factors in the same way as each other, so a unified approach must also allow for differences, and for individual variations in cognition.

Conscious and unconscious judgement: do examiners follow the rules?

It is an acknowledged aspect of examination marking that examiners become faster in their decision-making as time progresses during a session, whether it is a positive acknowledgement like that of the Team Leader on History 2, or a pejorative one, in the

controls placed on the speed of marking by means of the electronic system, described by the Principal Examiners in History 1. The idea that it becomes more 'intuitive' (TL, H2) sits well with the concept of a migration of conscious judgements from 'System 2' to the more automatic, instinctive judgements of System 1. So too does the decreased use of the mark scheme, for reference or in terms of the language which markers used, established in Chapter 6 and supported further by the findings of the questionnaire.

Kahneman and Frederick (2002) argued not only for the migration of judgements but also for the monitoring of the instinctive System 1, by the conscious and considered System 2. Their evidence suggested that an intuitive judgement was only expressed if it was endorsed by System 2. Much of the talk by senior examiners in the training meetings had the effect of making examiners conscious of factors which might affect their judgement, such as the length of an answer, or the discourse markers within it, and making explicit the circumstances under which it might be an appropriate signal for judgement. Examiners have a high level of awareness of their reactions to scripts, and make it clear that those reactions are monitored for inappropriate influence on judgements. This suggests that there is a degree to which conscious thought is put into monitoring intuitive judgement, to ensure that it is not inappropriate, although that is not the same as suggesting that the conscious thought is aimed at rule-based decision-making. There is some evidence, therefore, both for the migration of processes and for the monitoring of System 1 by System 2. However, it is also worth noting that the second Principal Examiner on History 1 warned her examiners that the time at which they speed up, later in the process, is when 'errors creep in'; a suggestion that intuitive judgements are not always monitored successfully.

There is a tacit acknowledgement even by the system itself that the mark scheme is not in fact the centre of the judgement process, despite the formal briefing for examiners; in fact scripts are placed at the centre of everything which an examiner must do throughout the training day and the marking period. The mark scheme is not formally introduced, and there is a great deal of emphasis on other factors and characteristics which are not part of the mark scheme; that emphasis is not a warning against using them as the basis of decisions, but rather a warning against *misusing* them as the basis of decisions, echoing the monitoring described above. The production of suggested stimuli to judgement, accompanied by caveats, is a characteristic feature of all training; a principle of 'productive uncertainty' is suggested below.

It is clear, though, that the process of decision-making is not a rule-based, standardised one, or at least not one in which the rules remain the same from script to script. Examiners take whichever tools are most useful to them in any given situation; there is certainly evidence to suggest that it is when decisions become difficult that they reach for the mark scheme, or for the sample scripts, or for a comparison between the sum of the marks they have awarded and their holistic, overall impression.

There is a strong level of self-awareness demonstrated by examiners, something which was particularly emphasised in the survey reported in Chapter 9. They will match the standard of their judgement to the standard of the principal examiner, as they perceive it. Sometimes this is done in rule-based ways, such as the systematic self-adjustment which was demonstrated in the qualitative data and substantially supported in the questionnaire response. Sometimes it is not a systematic adjustment, but the selection of a specific factor which is the most useful for the individual in predicting the correct judgement for scripts; that is the factor on which they can make a decision, which is most closely aligned to the principal examiner's judgement. Parrot's 'information' is one such factor; others are suggested – often taken from the focus of the exam – by senior markers. This calibration of the level of judgement is a personal activity, which is aided by the activities of the training meeting. The mark scheme is one tool which enables this; it is possible that the definitions of the terms are often left vague deliberately, to allow each individual to put a meaning on them which is most relevant to their judgement. That is to say, since their understanding of what each level is in terms of a script is based on analysis which is ultimately individual to them, the understanding of the meaning of the term in the mark scheme should also be left for that individual interpretation. Defining precisely what 'limited analysis' is in terms of the mark scheme might pin that down, but would not necessarily aid the marker in linking it to the scripts; indeed, it could disconnect their instinctual understanding of it by providing too specific a definition. This is perhaps an instance in which the Awarding Bodies rely on the guild knowledge of their examiners, on their construct of what 'analysis' actually is.

This is not, however, in sympathy with the overall impression of the training, which is that it primarily operates on the basis of making the thought processes of examiners less automatic, and to bring their conscious attention to their automatic judgements, essentially with the intention of reducing bias. It was suggested in Chapter 2 that the natural qualitative judgement process might be restricted by the analytical system which is used, particularly if examiners were not aware of the full extent of the criteria under which they make holistic judgements, and were unable to separate them as a result. This might be, in fact, a benefit,

in terms of preventing automatic judgements, given the concern of the meetings to draw examiners' attention to the signals which essays contain. Similarly the dissociation of the levels of the mark scheme from the grades with which teachers are familiar, and which they are used to using, prevents too-easy judgements which are not parallel with the required standard. But the balance is delicate, as demonstrated by Crisp's (2008a) finding that over-deliberated decisions were less likely to be aligned to the judgement of the principal examiner.

Although the marking system is designed to be standardised, to eliminate as much subjectivity as possible, ultimately the cognition of an individual examiner is not the same as any other, and each must rely on the tools which suit them best. It is only the subjective judgement of the examiner which is able to compensate and adjust so that the numerical label which each individual attaches to their own standard is aligned with that of the principal examiner.

If judgement is highly individual, and uses a variety of mechanisms and tools depending on the most useful, or cognitively economical, in the specific situation, there is one tool in particular which is very widely used. Comparison is a major principle of the way in which examiners make their decisions, with a variety of comparative material in use. It is used both unconsciously and consciously, particularly at a time of difficulty. It is perhaps because of this apparently universal mechanism ('all judgments are comparisons' (Laming, 2004)) that the use of anchor scripts to fine-tune decisions is more prevalent than the use of the mark scheme. It is reassuring that the data demonstrated care in the choice of these scripts on the part of principal examiners, in a way which spoke to the prevention of bias by several of the unconscious cognitive heuristics, and an emphasis on them during the training. Perhaps the system of examination marking by paired comparison (Pollitt, 2010) would be a more appropriate system for human markers, given the substantial evidence for the use of comparison in this study.

Conceptualisations: mental tools

It has been argued that annotation of examination scripts enables the creation of a 'visual map' of the answer (Crisp & Johnson, 2007), an idea which draws on reading comprehension theory to explain understanding of complex texts. Annotation of scripts is becoming less used as examination marking becomes an online process, and it was not seen to be used at all in this study. At several points in this thesis I have suggested that examiners are creating a 'mental map' by the use of analysis, or descriptive commentary; the senior examiners might

create an outline by supplying the kind of discourse markers a text lacked, for example, or examiners might pick up on specific terms used by the candidate during their read through to create an *aide-mémoire* of the content of the script. Instead of creating a 'visual map', examiners create an 'auditory map', either by speaking out loud or to themselves within the confines of their own head. This could also contribute to an extended working memory for examiners, if we accept Baddeley & Hitch's three independent components of working memory: rather than the visuo-spatial 'scratch pad' having to contain all relevant information, the capacity of the working memory could effectively be doubled by the use of the 'phonological loop' (Baddeley & Hitch, 1974). The auditory conceptual map might be held separate from, for example, the visually-stored mark scheme.

Beyond this mental map of the answer is the larger mental framework, suggested by Baird (2000) and by Wiliam's (1998) 'construct-referenced' assessment, which is initially populated by 'prototypes' (Crisp, 2010a) before further concrete examples are slotted into position, so that essay marking becomes 'one long discourse' (Vaughan, 1991). The extensive use of comparison, noted above, with anchor scripts, actual scripts and imaginary scripts, supports this concept. It is also a metaphor within which the cognitive heuristics easily sit. Although they rely, as Kahneman & Frederick noted, on using a more easily available alternative property in place of the property which is the more correct prompt for judgement (2002), they ultimately require relative values to govern judgement, particularly in terms of the *availability* heuristic and its dependence on the range of examples the mind can access.

Productive Uncertainty

The term 'productive uncertainty' has been suggested as a label for the way in which trainers deliberately refuse to make categorical statements or to lay down binding rules about aspects of judgement. The final category in the typology of comments in Chapter 5 draws on this, in its preservation of uncertainty as to whether a signal is true or false. Signals which examiners might rely on, and which are noted as potential indicators of quality, are accompanied by caveats about candidates' knowing what is required and making their essays appear to do so, without actually doing so. The emphasis may be on reducing the amount of variability, but it is variability of outcome which is controlled, not the variability of the individual's cognition, either between or within examiners.

The principle of productive uncertainty is present in the lack of definition of, or introduction to, the mark scheme. It is part of the attempts to make examiners consider

their own thinking, and to make problematic the factors to which their attention is naturally drawn. As suggested above, the training process must prevent examiners from making completely automatic judgments based entirely on their own past experience, which are not allied to the standard, because if judgements are entirely unconscious, then they cannot be adjusted. The examination marking process is made alien from normal classroom assessment, and from the grade-based outcomes of the process; it operates by 'making the familiar strange'. Nevertheless, it must also be 'productive.' Too much uncertainty leads to over-deliberation, to too great a need for time, and ultimately to inaccurate decisions. Some guidelines are given and a framework is created by the system to contain the subjective judgements of examiners.

This is not a principle which applies to the short answer question marking of the type which can rely on matching or pattern recognition. It is one which applies to the types of judgement which operate under the kind of constraints described in the introduction to this thesis, and the opening of Chapter 1: complex cognitive tasks whose requirements for attention and working memory would overpower the limitations of Simon's 'bounded rationality' (1992). It is the kind of judgement which does require, to some extent at least, an expert decision-maker: someone who is cognisant of the forms of guild knowledge and of the context within which the system sits, but whom the system can to some extent detach from their personal constructs of levels of quality, to reshape them in the form of the levels of the mark scheme.

Summary and implications for further research

The review of literature reflected on the fact that Verbal Protocol Analyses tended, by their nature, to be carried out with small samples, and that their strength was in the collective findings of many studies. This thesis has contributed to this aggregation, and in concerning itself specifically with essay questions, has extended the range of the limited number of studies on examiners' cognitive processes. It has provided empirical evidence for the extensive use of comparison by examiners, which has not been established as one of their primary decision-making behaviours previously, although some have assumed it. It has demonstrated that the process is far from uniform, despite the acceptably reliable outcomes, and I have suggested that this lack of uniformity is a symptom of a subjective judgement, for which the lack of categorical guidance provides a necessary cognitive environment: the 'productive uncertainty' principle. Some of the specific conclusions of the

study were validated in the questionnaire reported in Chapter 9, which also demonstrated that examiners do not always know what they are doing.

This study has had limitations: the very small sample of 'think aloud' participants was far from ideal. Most of the data is drawn from the very specific situation of the training meeting, in which decision-making is not quite the same process as individual marking, given the potential need to conform to social norms, although I would argue that the decisions are still those of individuals, which simply take place in a social setting. The training meeting is in some ways a naturally occurring version of the scenarios which some researchers have set up to investigate markers' thinking. Indeed the fact that this was not a simulation study, and that it could investigate some of the theoretical mechanisms of decision-making in the field was a strength.

This study has covered a wide range of theoretical concepts within examiner decision-making, in part because of the lack of previous research in the area of essay assessment. It was not possible to make the assumption that, for example, anchoring and adjustment would be present, and to investigate just that aspect of decision-making. What has been established is that there is a very wide range of processes which are going on inside the black box. Some of the factors which contribute to decision-making are not those which are considered to be appropriate (as demonstrated by some of the contradictory answers given by respondents to the questionnaire). The range of processes and factors bear further investigation, now that a range of exploratory synoptic studies, such as those reviewed in Chapter 2, and including the present one, have been carried out. Comparison in particular is a principle of assessment which has not been fully explored in relation to examiners' marking, as opposed to as a tool for the maintenance of standards, or for boundary setting. It may be that previous research which has been carried out in the context of simulation studies may simply not have had enough material available to examiners for comparison to be an effective tool for them. Further use of the 'think aloud' methodology might suggest further principles for the selection of anchor material, or the extent to which imaginary scripts are a problem for the system, as Pollitt (2010) suggested they might be.

The affect reactions of examiners and their projections of candidates which was briefly described in Chapter 8, and which I have explored in greater detail elsewhere (Elliott, 2010) would also be an interesting topic of further research. Elliott (2010) explored the potential tension between the roles of 'teacher' and 'examiner' which was also mentioned in Chapter 4 in relation to the establishment of a professional attitude. Some additional data in relation to this was gathered from responses to the open questions in the survey, which

were not included in Chapter 9 as they were not relevant to the topic of the thesis. A further qualitative interview study to explore the relationship between the two personas of teacher-examiners and how they reconcile the two roles would be of interest; in terms of relevance it might also help understanding of the problem of recruitment and retention of examiners which Awarding Bodies face, especially at a time when a move to online marking is likely to prompt the resignation of the most experienced markers, (a fear expressed in the responses to the open questions in the questionnaire).

Part of the impetus for this research was provided by the increasing debate over the use of computer programmes to assess students' work; that is, an entirely automated assessment process. Such systems have gained ground outside the UK: they are already in use in tests of English (Way & de Jong, 2010) and to some extent in the marking of SAT essays in the USA, in which an automated assessment is one marker in a double marked system. In the UK the reaction to such programmes has been negative: the most common example given is the speech by Winston Churchill which was given a failing grade by a computer programme.

Such programmes are 'trained' on hundreds of scripts marked by expert human examiners; they extrapolate the key criteria from the judgements they have been given, and mark further scripts using an algorithm based on that extrapolation. The objection has been that these programmes work on proxies, whereas with human examiners we 'know' what the basis of the judgement is: the rational application of the mark scheme and its criteria. I do not particularly support the use of computer aided assessment, but it felt appropriate that the debate should be based on a clear understanding of what was happening in the decision-making process, rather than an assumption that the system worked in the way it was nominally described.

It is clear that human examiners occasionally use proxies too; at times most of them use criteria other than those explicitly delineated in the mark scheme. Even the system, relying as it does on analytical marking based on sub-divided criteria, may be accused of using proxies which do not necessarily represent the qualities that are desirable holistically. The power of the human, subjective judgement is that it may reconcile the holistic and the analytical approaches by comparison, to ensure that assessment does not become invalid by decomposing the qualitative decision. A theme which recurs again and again throughout this thesis is the ability of the marker to adjust and to compensate for factors which skew judgement, or to adjudicate the situation where the lack of success in a presumed critical area has not in fact damaged the overall essay to the extent it might, and to reflect that in

the final judgement. But this very human recompense is not available to a computer system. It may be that the judgement of examiners is subjective. It is certainly not unvarying, either within or between markers. But that subjectivity is in itself a strength, which enables a fair, accurate and professional judgement to be made by an extraordinarily flexible and capable tool: the examiner.

References

Adelswärd, V. (1989) 'Laughter and dialogue: the social significance of laughter in institutional discourse', *Nordic Journal of Linguistics*, 12 (2), pp. 107–136

AQA (2007) *GCE English Literature B (2745) 2009 onwards* (Manchester: AQA). Available online < http://store.aqa.org.uk/qual/gce/pdf/AQA-2745-W-SP-10.PDF> (Accessed 12th December 2010)

AQA (2009) *General Certificate in Education AS History 5041 Alternative B Unit 1 Mark Scheme 2009 examination – January series* [Online] (AQA). Available from <http://store.aqa.org.uk/qual/gceasa/qp-ms/AQA-HS1B-W-MS-JAN09.PDF> (Accessed 13th July 2009)

Baddeley, A.D. & Hitch, G.J. (1974) 'Working memory', in G.H. Bower (ed.) *The psychology of learning and motivation* (Vol. 8) (London: Academic Press)

Baird, J.-A. (1996) *What's in a name? Experiments with blind marking in A level Examinations,* A paper presented at the British Psychological Society Conference in London on 17 and 18 December.

Baird, J.-A., (2000) 'Are examination standards all in the head? Experiments with examiners' judgements of standards in A level examinations', *Research in Education*, 64, pp. 91–100

Baird, J. and Scharaschkin, A. (2002) 'Is the Whole Worth More than the Sum of the Parts? Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-Level Examination Performances', *Educational Studies*, 28(2), pp.143-–62

Barritt, L., Stock, P.L. & Clark, F. (1986) 'Researching practice: evaluating assessment essays', *College Composition and Communication*, 37(3), pp.315–327

BERA (2004) *Revised ethical guidelines for educational research* (British Educational Research Association; London)

Berthoz, A. (2003) *Emotion and reason the cognitive neuroscience of decision making*. Translated by Giselle Weiss. (Oxford: OUP)

Bettenhausen, K., & Murnighan, J. K. (1985) 'The emergence of norms in competitive decision-making groups', *Administrative Science Quarterly*, 30(3), pp. 350–372

Black, B., & Bramley, T. (2008), 'Investigating a judgemental rank-ordering method for maintaining standards in the UK examinations', *Research Papers in Education*, 23(3), pp. 357–373

Bolger, F., & Wright, G. (1992) 'Reliability and validity in expert judgement', in Wright, G., & Bolger, F. (eds.) *Expertise and decision support* (New York: Plenum), pp. 47–76

Bramley, T. (2007) 'Paired comparison methods', in: P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.) *Techniques for monitoring the comparability of examination standards* (London: QCA)

Britton, J. (1964) *The multiple marking of compositions* (London: HMSO)

Britton, J. & Martin, N. (1989) 'English teaching – is it a profession?', *English and Education*, 23(2), pp. 1–8

Broadbent, D. E. (1982) 'Task combination and selective intake of information', *Acta Psychologica*, 50, pp. 253–290

Broadfoot, P., & Black, P. (2004) 'Redefining assessment? The first ten years of Assessment in Education', *Assessment in Education*, 11(1), pp.

Camerer, C.F., & Johnson, E.J. (1991) 'The process-performance paradox in expert judgement: how can the experts know so much and predict so badly?', in Ericsson, K.A., & Smith, J. (eds.) *Towards a general theory of expertise: prospects and limits* (Cambridge: Cambridge University Press), pp. 195–217

Chambers (1998) *The Chambers Dictionary* (Edinburgh: Chambers)

Chase, W.G., & Simon, H.A. (1973) 'The mind's eye in chess', in Chase, W.G. (ed.) *Visual information processing* (New York: Academic Press), pp. 215–281

Christie, T., & Forrest, G.M. (1981) *Defining Public Examination Standards*, Schools Council Research Studies (London: Macmillan Education)

Coffman, W.E., (1966) 'On the validity of essay tests of achievement', *Journal of Educational Measurement*, 3(2), pp. 151–156

Cox, R. (1967) 'Examinations and higher education: a survey of the literature', *Higher Education Quarterly*, 21(3), pp. 292–349

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

Crisp, V. (2008b) 'The validity of using verbal protocol analysis to investigate the processes involved in examination marking', *Research in Education,* 79(1), pp. 1–12

Crisp, V. (2010a) 'Judging the grade: exploring the judgement processes involved in examination grading decisions', *Evaluation & Research in Education*, 23(1), pp. 19–35

Crisp, V. (2010b) 'Towards a model of the judgement processes involved in examination marking', *Oxford Review of Education*, 36(1), pp. 1–21

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

Daugherty, R. (1988) *Examining Geography at 16+ A study of decision-making in two geography examinations* (London: Secondary Examinations Council)

Davis, J.H. (1996) 'Group decision making and quantitative judgements: a consensus model', in White, E., & Davis, J.H. (eds.) *Understanding group behaviour: consensual action by small groups*, vol. 1 (Mahwah, NJ: Erlbaum), pp. 35–59

Davis, J.H., Stasson, M.F., Parks, C.D., Hulbert L., Kameda, T., Zimmerman, S.K., & Ono, K. (1993) 'Quantitative decisions by groups and individuals: voting procedures and monetary awards by mock civil juries', *Journal of Experimental Social Psychology*, 57(6), pp. 1000–12

Delaney, C.M. (2005, September) *The evaluation of university students' written work*, paper presented at the British Educational Research Association Annual Conference, Glamorgan. Available from <http://www.leeds.ac.uk/educol/documents/149754.doc> (Accessed 13th November 2010)

DfES (2004) *Pedagogy and Practice: Questioning* (London: HMSO)

Djakow, Petrowski & Rudik (1927) *Psychologie des Schachspiels* [*Psychology of chess*] (Berlin: Walter de Gruyter)

Dux, P.E., Tombu, M.N., Harrison, S., Rogers, B.P., Tong, F. & Marois, R. (2009) 'Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex', *Neuron*, 63(1), pp. 127–138

Edexcel (2005) *Edexcel Advanced Subsidiary GCE in English Literature (8180) Edexcel Advanced GCE in English Literature (9180) Issue 5 Specification* [Online] (Edexcel). Available from <http://www.edexcel.com/migrationdocuments/GCE%20Curriculum%202000/271931_English_Lit_Iss5.pdf> (Accessed 13th July 2009)

Edexcel (2010) *Assessment Associates Survey 2010*

Edgeworth, , F.Y. (1890) 'The element of chance in competitive examination', *Journal of the Royal Statistics Society*, 53, pp.460–475.

Elliott, V.F. (2010) 'Judging a candidate or a script? Affect reactions and empathetic projections in examiners of A level History and English'. Paper presented at the Association for Educational Assessment Europe's 11th Annual Conference, Oslo 2010.

Ericsson, K.A. (2006) 'An introduction to *Cambridge Handbook of Expertise and Expert Performance*: its development, organization and content', in Ericsson, K.A., Charness, N., Feltovich, P.J., & Hoffman, R.R. (eds.) *Cambridge Handbook of Expertise and Expert Performance* (Cambridge University Press: Cambridge), pp. 3–19

Ericsson, K.A., Charness, N., Feltovich, P.J., & Hoffman, R.R. (2006) *Cambridge Handbook of Expertise and Expert Performance* (Cambridge University Press: Cambridge)

Ericsson, K.A. & Lehmann, A.C. (1996) 'Expert and exceptional performance: evidence of maximal adaptation to task constraints', *Annual Review of Psychology*, 47, pp. 273–305

Ericsson, K.A. & Kintsch, W. (1995) 'Long term working memory', *Psychological Review*, 102(2), pp. 211–245

Ericsson, K.A., Krampe, R.T., & Tesch-Römer, C. (1993), 'The role of deliberate practice in the acquisition of expert performance', *Psychological Review*, 100, pp. 363–406

Ericsson, K. A. & Simon, H. A. (1980) 'Verbal reports as data'**,** *Psychological Review***,** 87, pp. 215–251

Ericsson. K.A. & Simon, H.A. (1993) *Protocol Analysis: Verbal Reports as Data* (London: MIT Press)

Eysenck, M.W. & Keane, M.T. (2000) *Cognitive Psychology* (Hove: Psychology Press)

Finucane, M.L., Alhakami,A., Slovic, P. & Johnson, S.M. (2000) 'The affect heuristic in judgements of risks and benefits', *Journal of Behavioral Decision Making*, 13, pp. 1–17

Fiske, S.T. & Taylor, S.E. (1984) *Social Cognition* (California; Addison-Wesley Publishing Company)

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

Gilovich, T., Griffin, D. & Kahneman, D. (2002) *Heuristics and Biases: the Psychology of Intuitive Judgement* (Cambridge: CUP)

Graesser, A.C., Mills, K.K. & Zwaan, R.A. (1997) Discourse comprehension, *The Annual Review of Psychology*, 48, pp. 163–189

Greatorex, J., & Bell, J. (2004) 'Does the gender of examiners influence their marking?', *Research in Education*, 71, pp. 25–36

Greatorex, J. & Bell, J.F. (2008) 'What makes AS marking reliable? An experiment with some stages from the standardisation process', *Research Papers in Education*, 23(3), pp. 333–355

Green, A. (1998) *Verbal protocol analysis in language testing research a handbook* (Cambridge: CUP)

Hamp-Lyons, L. (1991) 'Reconstructing "Academic Writing Proficiency"', in L. Hamp-Lyons (ed.) *Assessing second language writing in academic contexts* (Norwood, NJ: Ablex Publishing Corps), pp. 127–153

Harris, R. (2001) 'Why essay writing remains central to learning history at AS level', *Teaching History*, 103, pp. 13–16

Hartog, P. & Rhodes, E.C. (1935) *An examination of examinations: being a summary of investigations on the comparison of marks allotted to examination scripts by independent examiners* (London: Macmillan)

Hollingshead, A.B., McGrath, J.E., & O'Connor, K.M. (1993) 'Group task performance and communication technology: a longitudinal study of computer-mediated versus face-to-face work in groups', *Small group research*, 24(3), pp. 307–33

Jackson, B. (1965) *English versus examinations a handbook for English teachers* (London: Chatto & Windus)

Jacowitz, K.E, & Kahneman, D. (1995) 'Measures of Anchoring in Estimation Tasks', *Personality and Social Psychology Bulletin*, 21(11), pp. 1161–6

Johnson, M. & Shaw, S. (2010) 'Towards an understanding of the impact of annotations on returned examination scripts', *Research Matters*, 10, pp. 16–21

Kahneman, D. & Frederick, S. (2002) 'Representativeness revisited: attribute substitution in intuitive judgement', in Gilovich, T., Griffin, D. & Kahneman, D. *Heuristics and Biases: the Psychology of Intuitive Judgement* (Cambridge: CUP), pp. 49–81

Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgement under uncertainty: heuristics and biases*, (Cambridge: CUP)

Kerr, N.L., MacCoun, R., & Kramer, G.P. (1996) 'Bias in judgment: comparing individuals and groups', *Psychological Review*, 103, pp. 687–719

Kerr, N.L., Niedermeier, K.E., & Kaplan, M.F. (1999) 'Bias in jurors vs bias in juries: new evidence from the SDS perspective', *Organizational Behavior and Human Decision Processes*, 80(1), pp. 70–86

Kerr, N.L., & Tindale, R.S. (2004) 'Group performance and decision making', *Annual Review of Psychology*, 55, pp. 623–55

Kimber, P (1984) a talk given by Peter Kimber, Scottish Examination Board, at the Secondary Examinations Council, 14.7.84.

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

Lavoie, P. & Grondin, S. (2004) 'Information processing limitations as revealed by temporal discrimination', *Brain and Cognition*, 54, pp. 198–200

Marois, R. & Ivanoff, J. (2005) 'Capacity limits of information processing in the brain', *Trends in Cognitive Sciences*, 9(6), pp. 296–305

Marshall, B. (2001) 'Marking the essay: teachers subject philosophies as related to their assessment', *English in Education*, 35(3) pp. 42–57

Marshall, B. (2011) *Testing English: Formative and Summative Approaches to English Assessment* (London: Continuum)

Massey, A. & Foulkes, J. (1994) 'Audit of the 1993 KS3 Science National Test Pilot and the concept of quasi-reconciliation', *Evaluation and Research in Education*, 9(3), pp. 119–132

Merriam- Webster (1979) *Webster's New Collegiate Dictionary* (Springfield, MA: G. & C. Merriam Company)

Milanovic, M. & Saville, N. (1996) 'A study of the decision-making behaviour of composition markers', in Milanovic, M. & Saville, N. (eds.) *Performance Testing, Cognition and Assessment selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnheim* (Cambridge: CUP), pp. 92–114

Murphy, R. (1979) 'Removing the marks from examination scripts before remarking them: does it make any difference?' *British Journal of Educational Psychology*, 49, pp. 73–78

Nádas, R., & Suto., I. (2009) 'Speed isn't everything: a study of examination marking', *Educational Studies*, 36(1), pp. 115–118

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

OfQual (2010) *GCSE, GCE, principal learning and project code of practice* (Coventry: Qualifications and Curriculum Authority)

Partington, J. (1994) 'Double-marking students' work', *Assessment & Evaluation in Higher Education*, 19(1), pp. 57–60

Pike, G.R. (1999) 'The constant error of the halo in educational outcomes research', *Research in Higher Education*, 40 (1), pp. 61–86

Poldrack, RA. (2000) 'Imaging brain plasticity: conceptual and methodological issues--a theoretical review', *Neuroimage*, 12, pp.1–13

Pollitt, A. (2004) *Let's stop marking exams*, paper presented at the IAEA Conference, Philadelphia

Pollitt, A. (2010) *How to assess writing reliably and validly*, paper presented at AEA-Europe, Oslo

Pollitt, A and Elliott, G (2003) *Finding a proper role for human judgement in the examination system,* Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', April 2003

Polyani, M. (1958) *Personal Knowledge: towards a post-critical philosophy* (Chicago: University of Chicago Press)

QCA (2006) *GCSE, GCE< VCE< GNVQ and AEA code of practice 2006/7* (London: Qualifications and Curriculum Authority)

Raikes, N., Fidler, J. & Gill, T. (2010) 'Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology, *Research Matters*, 10, pp. 21–7

Robbins, J.H. (2008) *Connoisseurship, coursework and the credibility of teacher assessments*, presentation at the Institute of Educational Assessors National Conference, 3rd May 2008. Available online at <http://www.ciea.org.uk/news_and_events/events_listing/past_events/iea_national_conference_2008.aspx> [Accessed June 6th 2011]

Robson, C. (2002) *Real World Research* (2nd ed.) (London: Wiley & sons)

Rosch, E. (1978) 'Principles of Categorization', in Rosch, E. & Lloyd, B.B. (eds), *Cognition and Categorization* (Hillsdale: Lawrence Erlbaum Associates), pp. 27–48

Sadler, D.R. (1987) *Defining and achieving comparability of assessments* (Brisbane: Board of Secondary School Studies)

Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems', *Instructional Science*, 18, pp.119–144

Sadler, D.R. (2009) 'Transforming holistic assessment and grading into a vehicle for complex learning', in G. Joughin (ed.) *Assessment, learning and judgement in higher education* (Dordrecht: Springer), pp. 45–63)

Sanderson, P. (2001) *Language and Differentiation in examining at A level*. Unpublished PhD thesis, University of Leeds.

Scottish Examining Board (1992) *Investigation into the effects of the characteristics of candidates and presenting centres on possible marker bias* (Scottish Examination Board internal report)

Sfard, A. (1998) 'On two metaphors for learning and the dangers of choosing just one', *Educational Researcher*, 27(2), pp. 4–13

Shanteau, J. & Stewart, T. R. (1992) 'Why study expert decision making? Some historical perspectives and comments', *Organizational Behavior and Human Decision Processes*, 53, pp. 95–106

Simon, H. (1992) *Economics, Bounded Rationality and the Cognitive Revolution* (Aldershot: Edward Elgar Publishing)

Slovic, P., Finucane, M., Peters, E., MacGregor, D.G. (2002) 'The affect heuristic', in Gilovich, T., Griffin, D., & Kahneman, D. (eds.) *Heuristics and biases: the psychology of intuitive judgement*. (Cambridge: Cambridge University Press)

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

Spelke, E.S., Hirst, W.C. & Neisser, U. (1976) 'Skills of divided attention', *Cognition*, 4, pp. 215–230

Stasser, G., & Titus, W. (1985) 'Pooling of unshared information in group decision making: Biased information sampling during discussion', *Journal of Personality and Social Psychology, 53,* pp.81–93.

Straus, M.A., & McGrath, J.E. (1994) 'Does the medium matter? The interaction of task type and technology on group performance and member reactions', *Journal of Applied Psychology*, 79(1), pp. 87–97

Suto, W.M.I., & Greatorex, J. (2006) 'What do GCSE examiners think of "Thinking Aloud"? Interesting Findings from a Preliminary Study', paper presented at the British Educational Research Association annual conference, University of Warwick

Suto, W.M.I., & Greatorex, J. (2008a) 'A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations', *Assessment in Education: Principles, Policy and Practice*, 15(1), pp. 73–89

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

Thompson, G., & Bailes, S. (1926) 'The reliability of essay marks', *Forum of Education*, 4, pp. 85–91

Thorndike, E.L. (1920) 'A constant error in psychological ratings', *Journal of Applied Psychology*, 4, pp. 469–477.

Tversky, A. & Kahneman, D. (1974) 'Judgement under uncertainty: heuristics and biases', *Science*, 185, pp. 1124–1131

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, NJ: Ablex Publishing Corps)

Voss, J.F., & Wiley, J. (2006) 'Expertise in History', in Ericsson, K.A., & Smith, J. (eds.) *Towards a general theory of expertise: prospects and limits* (Cambridge: Cambridge University Press), pp. 569–584

Walford, G. (2001) *Doing qualitative educational research: a personal guide to the research process* (London: Continuum)

Way, D., & de Jong, J. (2010) 'Auto-essay scoring'. Paper presented at *Pearson Standards and Assessment Conference*, 25-26 March, St Hugh's College, University of Oxford.

Weigle, S. (1998) 'Using FACETS to model rater training effects', *Language Testing* 15(2), pp. 263–87

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

Wertheimer, A. & Miller, F. G. (2007) Payment for research participation: a coercive offer?, *Journal of Medical Ethics*, 34, pp. 389–392

Wiliam, D. (1998) 'The validity of teachers' assessments', paper presented at the 22nd annual conference of the International Group for the Psychology of Mathematics Education, Stellenbosch, South Africa. Available from <http://learn.shorelineschools.org/spec/wasl/documents/wiliam_validity_of_teachers_assessments.pdf> [Accessed July 6th 2011]

Wilmut, J. (1984) *A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them*. AEB Research Report RAC 315.

Wineburg, S. (1991) 'Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence', *Journal of Educational Psychology*, 83, pp. 73–87

Wolf, A (1993) *Assessment Issues and Problems in a Criterion-based System* (London: Further Education Unit)

Yao Q., Qi Z. & Xi Z. (2008) 'A comparison of the quality of holistic and analytical marking of English compositions', *Foreign Languages Research*, 2008(5)

Zajonc, R.B. (1980) 'Feeling and thinking: preferences need no inferences', *American Psychologist*, 35(2), pp. 151–175

Appendix: Questionnaire

These two pages give a break down of the statements in section one of the survey. A facsimile copy of the questionnaire follows.

Comparison (7)

- "So I think it's 25. Let's check it against the sample scripts."
- "A better script would have thought about context too."
- "I use the sample scripts all the way through the marking session."
- "Not sure about this one. How does it compare to the sample scripts?"
- "Right, okay. So how does it compare to the sample scripts? It's better than M. Not as good as T. That would make it Band 4. So… 18 marks."
- "That makes 26. But it wasn't as good as the last one. Hmm. Make it 24."
- "This second essay is stronger than the first. Did I miss something?"

Quality of Written Communication (3)

- "The content is okay, but the style's a bit weak."
- "The style of writing makes it seem better than it is."
- "The quality of written communication is terrible – the content is much better than it appears."

Heuristics

Anchor and Adjustment (2):

- "I can see from the beginning this is going to be a top band answer. Where does it belong in the band though?"
- "It looked as if it was better than this to start, but it's basically one long point so I'm going to put it down a bit."

Representativeness (4 + 1 expected to prompt a 'no'):

- "Hmm. This is very short. I don't hold out much hope for this."
- "I have a kind of idea of what a band 3 looks like."
- "It looks short but the handwriting is very small. Very deceptive."
- "This just feels like a top level script."

- “I don’t need to read the rest of this. It’s clear from the first page it isn’t going to get any better.”

Unofficial heuristics (2):

- “I know we use bands, but I know where the grade boundary is and I use that as a reference point.”
- “My team leader told me that this factor was the most important in choosing a band.”

Affect (4)

- “At last! Someone who’s answered the question.”
- “I feel really sorry for this kid. If I give this a five it’s going to fail. But it’s only worth five.”
- “I think I can scrape this into the bottom of the next band up.”
- “Oh, why did you write that? You silly silly child.”

Self-adjustment (2) (Contrast pair)

- “I know that I tend to be harsh, so I need to give them the benefit of the doubt and bump it up a mark.”
- “I was too generous on the training scripts, so I need to choose the lower of the two marks.”

Deduction (4)

- “He’s working through a checklist he’s been given.”
- “Must be a girl with that handwriting.”
- “This candidate has been taught really badly.”
- “This candidate should have done better. There’s potential but you’re not marking that. You’re marking what’s there.”

Mark scheme/ marking behaviour (5)

- “It gets easier to make a decision. I mean, fifty scripts down the line, it’s more intuitive.”
- “By the time I’m marking the last scripts I don’t need to use the mark scheme at all.”

- “I have the mark scheme in front of me at all times.”
- “I’ve not usually got a firm figure in my mind the first time I read through…then the second time I make a decision.”
- “Is this some awareness? Or is it limited awareness?”